Computer Methods and Programs in Biomedicine 235 (2023) 107528

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Comparison of deep learning models for digital H&E staining from unpaired label-free multispectral microscopy images

Jesus Salido[a,*], Noelia Vallez[b], Lucía González-López[c], Oscar Deniz[b], Gloria Bueno[b]

[a] IEEAC Dept. (ESI-UCLM), Pº de la Universidad 4, Ciudad Real, 13071, Spain
[b] IEEAC Dept. (ETSII-UCLM), Avda. Camilo José Cela s/n, Ciudad Real, 13071, Spain
[c] Hospital Gral. Universitario de C.Real (HGUCR), C. Obispo Rafael Torija s/n, Ciudad Real, 13005, Spain

ARTICLE INFO

Article history:
Received 23 January 2023
Revised 27 March 2023
Accepted 3 April 2023

Keywords:
Virtual staining
Digital staining
Multispectral imaging
Digital pathology
GAN (Generative adversarial network)
Cycle consistency
Contrastive learning
Style transfer
Image quality assessment

ABSTRACT

Background and objective: This paper presents the quantitative comparison of three generative models of digital staining, also known as virtual staining, in H&E modality (i.e., Hematoxylin and Eosin) that are applied to 5 types of breast tissue. Moreover, a qualitative evaluation of the results achieved with the best model was carried out. This process is based on images of samples without staining captured by a multispectral microscope with previous dimensional reduction to three channels in the RGB range.

Methods: The models compared are based on *conditional GAN* (pix2pix) which uses images aligned with/without staining, and two models that do not require image alignment, *Cycle GAN* (cycleGAN) and *contrastive learning-based* model (CUT). These models are compared based on the structural similarity and chromatic discrepancy between samples with chemical staining and their corresponding ones with digital staining. The correspondence between images is achieved after the chemical staining images are subjected to digital unstaining by means of a model obtained to guarantee the cyclic consistency of the generative models.

Results: The comparison of the three models corroborates the visual evaluation of the results showing the superiority of cycleGAN both for its larger structural similarity with respect to chemical staining (mean value of SSIM $\sim$ 0.95) and lower chromatic discrepancy (10%). To this end, quantization and calculation of EMD (Earth Mover's Distance) between clusters is used. In addition, quality evaluation through subjective psychophysical tests with three experts was carried out to evaluate quality of the results with the best model (cycleGAN).

Conclusions: The results can be satisfactorily evaluated by metrics that use as reference image a chemically stained sample and the digital staining images of the reference sample with prior digital unstaining. These metrics demonstrate that generative staining models that guarantee cyclic consistency provide the closest results to chemical H&E staining that also is consistent with the result of qualitative evaluation by experts.

1. Introduction

Currently, the diagnosis of many diseases, including cancer, is carried out by direct microscopic observation of tissue samples by experienced pathologists. In its digitized version, a study is performed based on digital images (WSI, *Whole Slide Images*) obtained by scanned samples of tissue. The samples prepared from biopsies consist of very thin slices (2–5 μm) of the tissue under study, fixed on glass slides. To facilitate the observation of specific cells and tissue components, stains provide color and contrast discrimination to the translucent tissue components (i.e., cytoplasm, cell nucleus and other tissue structures). Such stains are applied in the delicate preparation process of the specimens. Histochemical staining techniques are a universal procedure for histopathology studies based on the observation of tissues on a microscopic scale.

There are numerous purpose-driven staining techniques [1,2], but among them, hematoxylin and eosin dye staining (henceforth H&E) accounts for approximately 80% of human tissue staining procedures globally. H&E staining used in primary histologic stud-

* Corresponding author.
E-mail address: jesus.salido@uclm.es (J. Salido).

https://doi.org/10.1016/j.cmpb.2023.107528

ies for cancer diagnosis is a standardized process that takes approximately 2 hours to complete. H&E staining provides tissues with a characteristic pink-bluish coloration, where purple bluish tones correspond to the cell nucleus, pink tones to the cytoplasm and cell matrix, and a mixture of both to the rest of the tissue structures. Other special stains are intended to reveal tissue-specific components (e.g., Periodic Acid Shiff - PAS), including at the molecular level (e.g., immunohystochemical staining - IHC), which help complete the histological study when necessary.

Despite the standardization of histopathology protocols, staining results produce staining and contrast variability between samples that derives from small changes in processing conditions (e.g., materials used, protocols followed in pathology laboratories, digital image acquisition, etc.). The aforementioned variability can cause inconsistency in the diagnosis offered by different pathologists and even reduce the performance of digital histopathology techniques (e.g., segmentation, classification, etc.). In addition, when special staining is required, the process is more laborious and costly, increasing the time required to final diagnosis. To alleviate the above limitations, attempts have been made in the last decade to modernize the histopathology workflow.

Virtual digital staining (hereafter referred to as *"digital staining"*) is a promising approach in digital image-based histopathological studies [3,4]. Digital staining refers to the use of computer algorithms to artificially recreate the staining effect in the digital image of a tissue sample, without physical manipulation of the sample. Some recent works propose the automatic computer generation of the image of a sample with different stains by means of a digital transformation between different staining modalities (*stain transformation* or *re-staining*) (e.g., H&E → PAS, H&E → IHC). Stain transformations can be chosen to highlight relevant histological features in the same sample without modifying the traditional sample preparation process [5,6]. Among the advantages of digital staining the following can be considered:

- *It is a conservative rather than destructive process*. The application of different staining modalities does not require new tissue and makes additional biopsies for alternative studies unnecessary.
- *It reduces time and costs*. The application of computerized image processing algorithms is almost immediate and does not require investment in dyes and other consumables needed in the physical preparation of the specimens.

Starting from a label-free sample that has not undergone the destructive transformation inherent in the staining process itself, additional advantages are obtained, such as those listed below [4]:

- Application of independent digital stains for the same tissue sample.
- Standardization and repeatability. Digital processing of virtual stains provides standardized results in terms of color and contrast, as it does not suffer from the variability of the physical process. Moreover, it always obtains the same result repeatedly applied to the same starting image.
- Digital blending of different staining modalities in specific regions of interest (ROI) from the same WSI.

To date, the problem of digital staining has been approached following different strategies:

1. Based on a reference image (i.e., at the pixel level). It is posed as a color transfer problem analogous to color normalization, which aims to reduce color variability between different samples belonging to the same staining domain [7]. The process consists of obtaining the color of each pixel of the target image from the color characteristics of a reference image, applying methods based on histogram equalization [8] or separation of constituent components of the staining color [9–14]. These methods start from a single, carefully chosen reference image. Consequently, they present difficulties in reflecting representative coloration of structures absent in the reference image.
2. Based on a reference distribution (i.e., at the structural level). This distribution represents the features, content, and style of a set of images, which are to be transferred to the input image. The feature transference must maintain the structural coherence of the input image once the transformation has been applied. To achieve these objectives, the methods are grouped into two categories:
   (a) *Style transfer*. This category includes methods that start from the separation of the content and style components of the image to carry out an optimization process that generates a new image combining both components in the desired proportion. Therefore, these methods propose a generative model that obtains a target image by combining the content and style components provided independently by each of the two input images. Both components are extracted independently by a pre-trained *Convolutional Neural Network* (CNN), hence called *Neuro Style Transfer* (NST). The generation is guided by minimizing the discrepancy of the content and style components between the input images and the generated image [15,16].
   (b) *Image-to-image translation*. These are methods that use *Generative Adversarial Networks* (GAN) [17] to simultaneously learn the content and style characteristics of two sets of images belonging to different domains to achieve the transformation of an image from one domain to the other. Thus, *adversarial loss* forces style transfer between images, while *cycle consistency* preserves the content between them. To obtain satisfactory results, some of these methods require a set of aligned image pairs between the two domains, as in the case of *conditional GANs* [18]. However, to obtain acceptable results when it is complex, or impossible to have pairs of aligned images between the domains involved in the transformation, solutions based on *Cycle GAN* [19–21] and *contrastive learning* [22] are proposed.
3. Strategies applied at the semantic level. While the previous two strategies are considered data-driven, this third category refers to task-driven methods (e.g., classification, segmentation). In this approach, an automated task is performed on the original images and the result of this is used for the conditional application of virtual staining [23].

The application of *Deep Neural Networks* (DNN) in medical image analysis over the last decade has boosted the development of digital pathology [24,25]. In the last five years, GANs have been used as a tool with great potential in tasks such as digital color normalization of histological digital images [7]. However, the application of these networks requires training the models with aligned (i.e., registered) image pairs. For this reason, in this context Cycle GAN has proven to be the most promising option for the application of digital staining systems, as they do not require aligned image pairs.

A GAN is a DNN whose output ($\hat{y}$) is a sample belonging to a class (Y), generated from learning the representative distribution of attributes of that class ($y \sim \text{pdata}(y)$). In a GAN there are two competing models: 1.- the generator (G), responsible for learning the attribute distribution of the target class (i.e., stained samples), and 2.- the discriminator (D), whose mission is to learn to differentiate a genuine specimen (y, i.e., chemical stained sample) versus one artificially generated by the first model ($\hat{y}$, i.e., digital stained sample). A GAN can be conditioned to generate an instance of the class, from an input (x, i.e., unstained sample) that shares certain attributes with the generated instance [18].

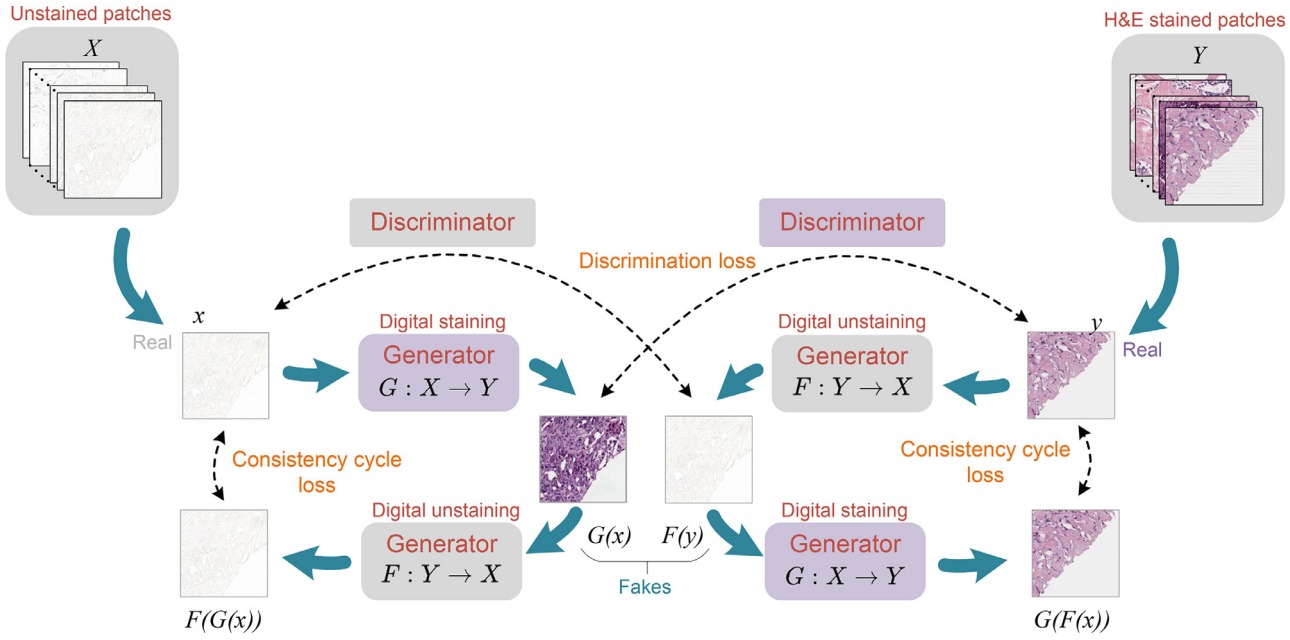

Fig. 1. Workflow of generative adversarial models with cyclic consistency.

In a Cycle GAN [19] (see Fig. 1), two domains or classes X and Y, are involved to learn a mapping of instances from one domain to the other ($G : X \to Y$, $F : Y \to X$). Thus, from an instance in either domain a new instance can be generated in the other domain (i.e., $G(x) = \hat{y}$, $F(y) = \hat{x}$). The generator-discriminator model scheme is duplicated in a Cycle GAN due to the double direction of generation that can be given from one domain to the other (i.e., $X \to Y$ and $Y \to X$). In addition, a new ingredient called *cycle consistency* appears, which guarantees that starting from an element in one domain, one can return to the same element by double generation (i.e.; $X \to Y \to X$ and $Y \to X \to Y$). That is, $F(G(x)) \sim x$, $G(F(y)) \sim y$. A salient feature of these networks is that their training does not require pairs of corresponding images in X and Y. The discriminative models consist of a CNN (e.g., VGG-16 [26]) and the generators consist of the layers of a convolutional encoder, ResNet [27] transform blocks, and layers of a deconvolutional decoder.

The application of contrastive learning to the problem of inter-domain image translation proposes a generative model based on patches to maintain content between portions whose mutual information is maximal to modify their appearance [22,28]. In these models, a CNN (e.g., ResNet) is used to obtain an internal representation of each patch in the *latent space* so that the generated patches are closer to the input patches compared to randomly generated patches. This type of model does not require aligned images between the domains to be transformed.

The evaluation of digital staining results is a crucial challenge, as it facilitates both the adjustment of the models used and the comparison of the results obtained by them. The *image quality assessment* (IQA) criteria traditionally used to evaluate digital staining models can be grouped into two categories, depending on the existence or absence of a reference image (i.e., chemically stained sample) with which to compare the image resulting from the process to be evaluated (i.e., digitally stained sample):

1. Computational analysis methods with reference image. These methods evaluate the structural and chromatic coherence between the reference image and the one obtained from the evaluated transformation process. In this case it is possible to define metrics based on the *human visual system* (HVS) that provide a measure of the fidelity of the distortion suffered by the reference image after a transformation process. Among the most commonly used metrics are SSIM (*Structural Similarity metric*), MS-SSIM (*Multi Scale Structural Similarity metric*) [29,30], VIF (*Visual Information Fidelity metric*) [31], FSIM (*Feature Similarity metric*) [32] and chromatic differences in the YCbCr color space at the pixel level. Recently, perceptual similarity metrics such as LPIPS (*Learned Perceptual Image Patch Similarity*) [33], based on the distance of activation vectors of deep feature maps obtained on CNNs trained for classification (e.g., AlexNet, VGG) [26,34], have been proposed. When using these metrics, the result of digital staining is considered a process that "distorts" the reference image constituted by the image of the same sample with chemical staining.
2. Task-based methods. These methods do not use a reference image. With them, the semantic consistency of the transformed image is evaluated following two strategies:
   (a) Evaluation of results in automatic tasks applied to digitally stained images (e.g., classification, segmentation, etc.) by comparing their results with those obtained on chemically stained images [24,35].
   (b) Blind visual evaluation by a panel of expert pathologists who perform an analysis of the images without knowing a priori the staining modality that the samples have undergone (i.e., chemical or digital) [6,36].

Regarding the evaluation of generative models, it is important to mention the FID (Fréchet Inception Distance) metric proposed to empirically estimate the degree of divergence between two sets of images (i.e., their distributions) using the feature space obtained in a DNN [37]. For this metric to provide representative values, it must be applied to datasets with a number of samples around 10 000.

1.1. Proposed work

This paper presents the comparison of three types of generative image-to-image translation models applied to H&E digital staining of label-free samples belonging to five subdomains, obtained by selecting three channels from multispectral microscopic images of breast tissue. The models compared are based on:

1. Conditional GAN network (pix2pix) trained with aligned input images [18];
2. Cycle GAN (cycleGAN) network trained with unpaired input images [19].
3. Generative network based on contrastive learning (CUT) trained with unpaired input images [22].

The comparison of the models is performed quantitatively by computational analysis methods to assess the structural similarity based on calculating the SSIM index. Since this method needs a reference image to compare the digital staining result, the inverse generative model $\hat{x} = F(y)$ is used, which guarantees cyclic coherence between the transformation domains X and Y. Thus, from chemically stained samples it is possible to obtain unstained samples to which the digital staining models to be compared are then applied. In our approach, instead of using the differences in brightness (Y) and chroma (Cb, Cr) obtained in YCbCr color space at the pixel level, a metric based on chromatic quantization and calculation of the color distance between chemically and digitally stained samples is used.

1.2. Main contributions

The main contributions of the proposed work can be summarized as follows:

- Use of a contrastive learning-based model to the digital staining problem.
- Comparison of three generative digital staining models (i.e., pix2pix, cycleGAN, CUT) applied to label-free samples acquired by multispectral microscopy.
- Comparison of digital staining models by chromatic quantification using clustering techniques and calculation of the color distance between clusters.

1.3. Related works

Although some studies do not attempt to digitally reproduce histochemical staining, augmented reality methods are proposed by superimposing masks (e.g., color or borders) on acquired digital tissue images. Therefore, they can be considered a non-conventional digital staining modality. For example, in the work of Litjens et al. [38] heat maps are generated overlaid on images of breast and prostate tissue with H&E staining to indicate the likelihood of cancer tissue to help the pathologist direct a more detailed examination. In other work, these masks even provide real-time feedback on the microscopic observation [39]. These works obtain results that can be considered a digital staining modality but are achieved with a very different approach from those that aim to digitally reproduce an effect analogous to that obtained with histochemical staining.

Table 1 summarizes information from deep learning based works related to the digital staining of unstained samples (i.e., label-free) and the transfer between different staining domains (also known as *digital re-staining*). In the analysis of these works, the aspects reflected in this table are: microscopy modality for sample image acquisition, types of tissues used, staining domains (source and destination), image training dataset with paired or unpaired images, architecture used and method of results evaluation.

The evaluation of the results determines the validity of the proposed solutions and paves the way for future improvements. The need for a training data set composed of aligned image pairs conditions the methods for evaluation of the results based on a reference image. Due to the very nature of the staining process, it is complex to obtain aligned image pairs, so the work has been approached using different strategies:

1. Application of automatic alignment (or image registration) algorithms for microscopy images [40–42]. These algorithms are valid when the discrepancies between images are small and can be corrected by applying geometric transformations to them. In general, the application of this strategy is discarded in the case of virtual staining in favor of the following ones.
2. Acquisition of images of unstained samples followed by chemical staining of the samples. Once virtual staining has been applied to the samples without staining (label-free samples), a set of aligned pairs of images (with chemical and digital staining) is obtained [43,44]. Several recent studies have utilized automatic alignment techniques to rectify misalignment resulting from the staining process, thereby augmenting the overall quality of the outcomes [45,46].
3. Inverse staining cycle. It is applied in trained models with cyclic consistency. In this case, chemically stained test samples are used, to which a digital unstaining is applied, followed by a digital staining process [47–49]. In the end, a set of aligned pairs of chemically and digitally stained images is obtained.

After analyzing related works, only one of them [50] has employed multispectral microscopy as a label-free modality to obtain specific virtual staining models on lung tissue using conditional GANs. That work does not present alternatives to the models and suffers from the need to start from aligned image pairs for the training of the models. With our proposal, we aim to further explore the possibilities of this label-free microscopy modality as a starting point for virtual staining models.

2. Materials and methods

2.1. Image acquisition

The biopsies used in this work were obtained thanks to the collaboration of the Hospital General Universitario de Ciudad Real (HGUCR), where the slides of a total of 13 pairs (with H&E and w/o staining) of breast biopsy samples were prepared. The H&E stained samples were digitized using a Leica Aperio CS2 scanner (www.leicabiosystems.com) with 20x and 40x magnification functions. These biopsies were classified by a pathologist as: malignant (n=8), normal (n=4) and benign (n=1). To capture the correspondent multispectral images (MSI) of the samples without staining, the IMA VIS hyperspectral microscope from Photon etc (www.photonetc.com) was used.

The microscope was equipped with three objectives 10x, 20x and 40x, a motorized stage in the X, Y, Z axes, and a programmable spectral filter with a resolution of 1 nm in a total range of 400–700 nm. The sample captures were performed with the 20x objective in the spectral range of 425–700 nm and a spectral step of 4 nm. The collaborator pathologist selected the regions of interest (ROI) in the WSI. That is, the most important areas for the study. The tissue within these ROIs was classified into 5 classes: *adipose* tissue, *non-tumorous* cellularity (besides pre-existing structures also include benign or pre-neoplastic lesions), *stroma, tumoral stroma* (consisting in pre-existing connective tissue and desmoplastic newly generated stroma), and *tumorous* (i.e., invasive cancer) tissue. The selected ROIs were decomposed in patches of two sizes $1\,024 \times 1\,024$ px and 512×512 px, to cover different tissue areas and obtain the final datasets for model training.

For the training of the digital staining models, RGB images of both chemically stained and non-stained samples were used in the 3 most informative channels selected from a previous dimensional reduction study carried out for the MSI. The composition of the two datasets used is indicated below in the Fig. 2. After the process, two datasets are produced:

Table 1
Comparative summary of related works.

Ref.	Micr. modality	Tissue	Orig. domain	Target domain	(**P**/**U**np)aired	DNN model	Assessment metric
[50]	hyperspectral (3 ch.)	lung	label-free	H&E	P	cond GAN	qual. (visual)
[54]	digital re-staining	colorectal	Ki67-CD8	FAP-CK	U	cycleGAN	quant. (segmentation results)
[43]	autofluorescence	salivary gland, thyroid kidney, liver, lung	label-free	H&E, Mason's trichrome + Jones	P	GAN	quant. (SSIM, YCbCr diff.)
[5]	digital re-staining	colorectal	H&E	CK18/CK19 (8 subdomains)	U	cond GAN, cycleGAN	qual. (perceptual study)
[36]	brightfield	carotid artery	label-free	H&E + PSR + Orcein	P	cond GAN (StarGAN)	qual. (blind visual eval.)
[47]	brightfield	prostate	label-free	H&E	P	cond GAN	quant. (SSIM, PCC, PSNR) qual. (blind visual eval.)
[4]	fluorescence (2 ch., DAPI + TxRed)	kidney	label-free	H&E + Masson's trichrome + Jones	P	cond GAN	quant. (SSIM, YCbCr diff.)
[48]	digital re-staining	breast, neuroendocrine	H&E	Ki67	U	cycleGAN (PC-StainGAN)	quant. (SSIM, MS-SSIM, PSNR, etc.)
[6]	fluorescence (2 ch., DAPI + TxRed) + digital re-staining	kidney	H&E	Masson's trichrome + Jones + PAS	P	GAN + cycleGAN	qual. (blind visual eval.)
[44]	autofluorescence + digital re-staining	kidney	H&E	PAS	P	GAN + cascade GAN	quant. (SSIM, YCbCr diff.)
[49]	optical coherence tomography (OCT)	coronary artery	label-free	H&E	U	Struct. constrained cycleGAN	quant. (PHV)
[45]	total absorption photoacustic remote sensing TA-PARS (3 ch.)	skin	label-free	H&E	P	cond GAN (pix2pix)	quant. (SSIM, RMSE) qual. (board visual eval.)
[46]	autofluorescence (DAPI + FITC + TxRed + Cy5)	breast	label-free	HER2	P	cond GAN (attention gated GAN)	quant. (blind scoring test, stat. dist. eval.)
Ours	multispectral (3 ch.)	breast	label-free	H&E	P + U	cond GAN + cycleGAN + CUT	quant. (SSIM, color distance) qual. (blind scoring test)

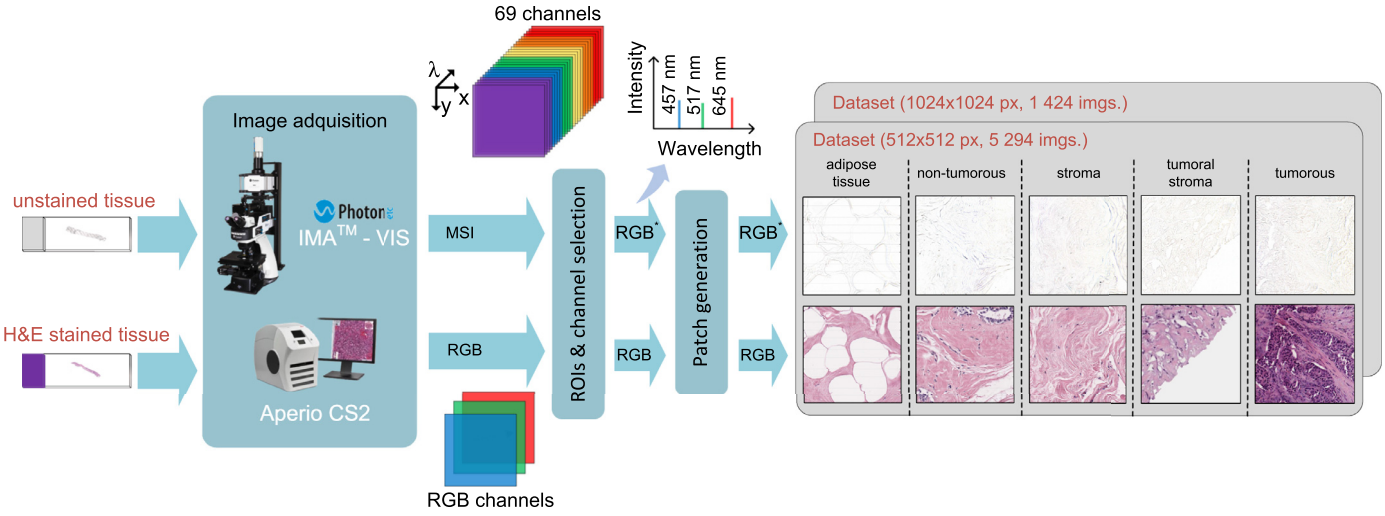

Fig. 2. Schematics for dataset acquisition.

1. Dataset of patches 1 024 × 1 024 px. Composed of 1 424 pairs of images (w/o staining of the same tissue area).
2. Dataset of patches 512 × 512 px. Composed of 5 294 pairs of images.

2.2. Models, implementation and training

The datasets mentioned in the previous section were used for the training of three digital staining models: (M1) pix2pix, (M2) cycleGAN and (M3) CUT. The implementation of the pix2pix model requires pairs of aligned images for training. However, because sample preparation processes (i.e., cutting, staining, etc.) cause changes in tissue structure, it was necessary to perform a computational registration process to correctly align the image pairs. To register the images, a 2D affine transformation implemented in MatLab (Mathworks) was utilized (refer to the `imregister` function) over the entire WSI. Then the resulting output was validated manually, and the regions of interest (ROIs) were cropped from the previously registered whole slide images (WSI). These patches were used to create the training dataset for the pix2pix model. In contrast, the mentioned registration was unnecessary for cycleGAN and CUT models, as they do not require image alignment.

The three digital staining models were derived on the base Pytorch implementations proposed by their respective authors, which are detailed in the cited references [18,19,22] (see also https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix and https://github.com/taesungp/contrastive-unpaired-translation). These references provide information on the architectures used and the ob-

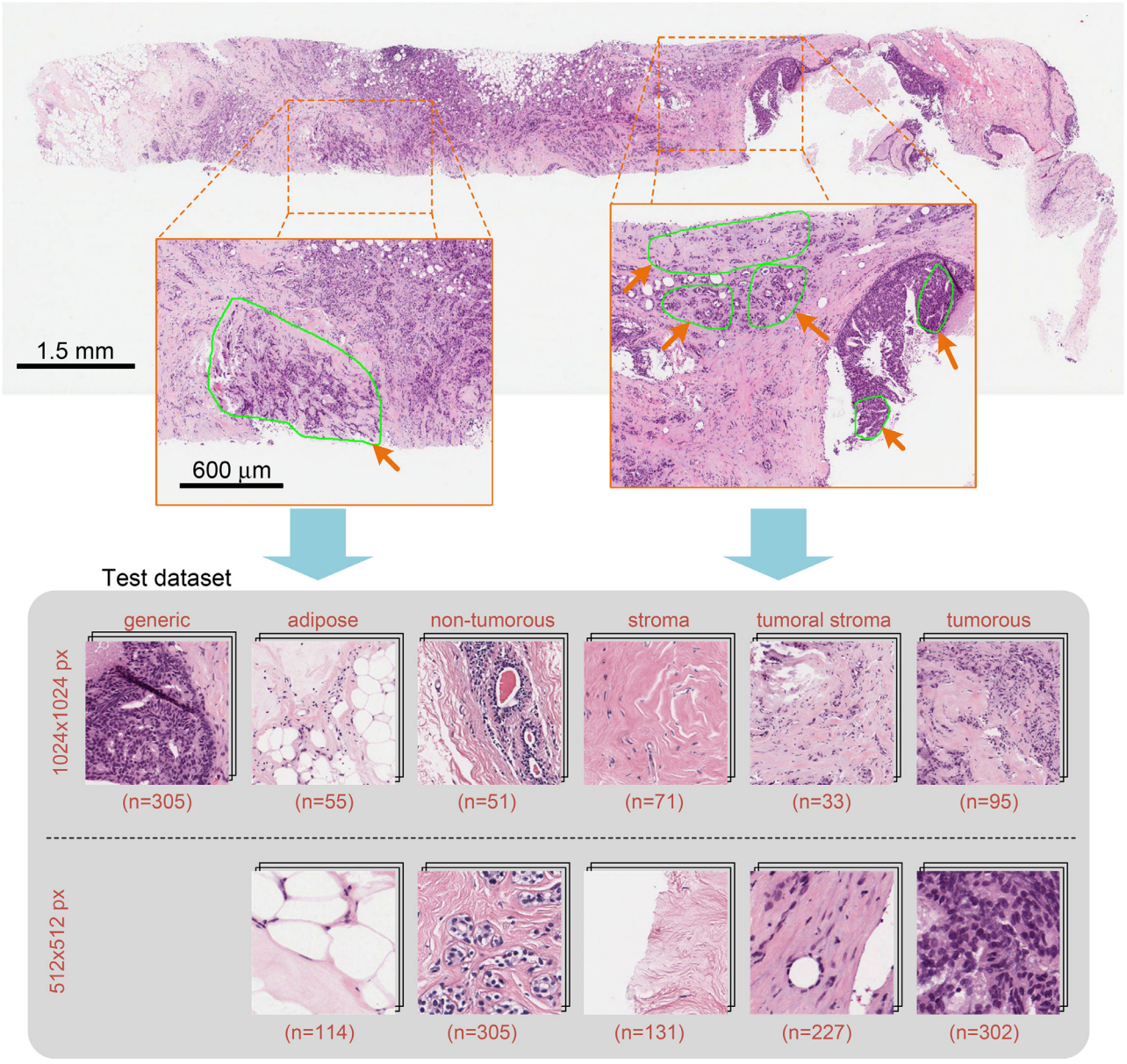

Fig. 3. Test dataset and test image subsets for models assessment.

jective loss functions used during their training. In particular, the batch size decision for pix2pix model and cycleGAN was made to satisfy the memory constraints of the hardware being used (i.e., Intel Xeon E5-2620 2 Ghz with a 8 GB GPU NVIDIA Quadro P400). In the experiments carried out, it was observed that this value ensured an adequate advance of the convergence during training. The training parameters for the different implementations of each model were: initial learning rate of 0.0002, batch size of 1, a number of epochs of 200, 400 and 600, and the re-scaling of images to $256 \times 256 \times 3$ px. This re-scaling was necessary to reduce the amount of memory required to train the models.

To start the experimentation, the three models were trained with images without differentiating the type of tissue. Thus, we obtain three generic models. Afterwards, with the intention of capturing the characteristics of each tissue and analyzing the functioning of more specific digital staining models, individual models of cycleGAN and CUT were trained for each of the 5 tissue types present in the datasets: (T1) adipose, (T2) non-tumorous, (T3) stroma, (T4) tumoral stroma, and (T5) tumorous.

2.3. Comparison of models for digital staining

Due to the very nature of histological sample preparation, it is not possible to achieve perfect alignment between unstained images and their corresponding images with chemical staining (see the samples of the input dataset in Fig. 4). Therefore, it is not possible to apply metrics based on computational analysis for the evaluation of image quality (IQA) applied to the sample with digital staining taking as reference the corresponding image with chemical staining. To overcome this difficulty, it was decided to evaluate the digital staining models by performing the digital staining of reference samples obtained by digital un-staining of test samples with chemical staining unseen before by the models. In this way, it is possible to compare the different digital staining models by comparing the results with perfectly aligned chemical staining reference images.

To generate the datasets of test images, a WSI not used during the training of the models was selected. In this sample, the expert pathologist labeled 19 ROIs (see input dataset in Fig. 3) from which

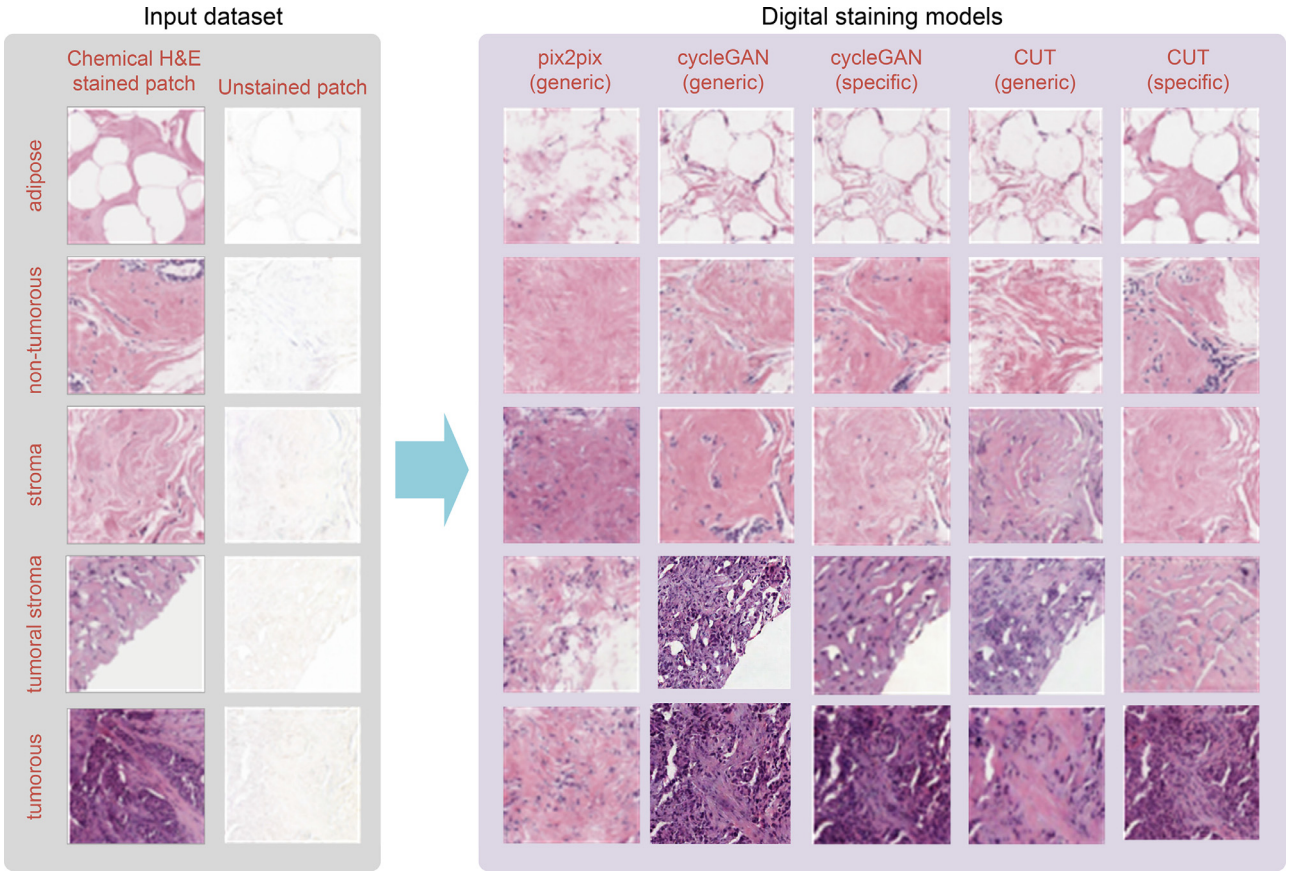

Fig. 4. Results of digital staining with the generic and specific (by tissue type) models from input images (target chemical staining and unstained images).

they extracted patches to make the test dataset to evaluate and compare the models.

The Fig. 3 summarizes the process of obtaining and composing the datasets for evaluation and comparison of both generic and specific digital staining models.

3. Results and discussion

Figure 4 shows the results obtained for generic (i.e., without distinction of tissue type) and specific (i.e., for each type of tissue) digital staining models, the latter for cycleGAN and CUT. The evaluation stage of the models is designed to answer the following questions:

- **(q1) Question 1**: Which of the digital staining models has the greatest structural similarity to chemical staining?
- **(q2) Question 2**: Which model obtains the least chromatic discrepancy with respect to chemical staining?
- **(q3) Question 3**: Does tissue-specific modelling improve digital staining outcomes?
- **(q4) Question 4**: Are digitally stained images comparable to chemically stained images and sufficient for diagnostic use?

The answer to these questions should be associated with statistical estimates of variability and significance of the results, to ensure that the conclusions obtained are general and not the result of chance. Next sections explain our approach to answer the posed questions.

3.1. (q1 answer) - Structural similarity assessment

To compare the models with respect to structure similarity, the subset of 305 generic test images (1 024 × 1 024 px) was used. For that subset, the SSIM metric [29,30] was calculated with the digital staining result achieved by each of three generic models, with the chemical staining taken as the reference image.

The metrics are calculated over the scaled version (256 × 256 px) of the patches taken as input to the models and outputs getting from them. Using non-rescaled patches of size 256x256 pixels was also considered, but the results did not show a significant improvement. One possible explanation is that reducing the area size for training the model leads to a loss of structural patterns in the tissue. Therefore, we concluded that subsampling improves the results compared to considering smaller patches, which maintain spatial resolution but lose the characteristic tissue patterns that provide contextual information and enable generalization.

The chosen subsampling method also affects the evaluation metrics, but since the reference images are also re-scaled, the calculations are equally affected in all the three compared models, and do not impact the conclusions drawn from the comparison. The SSIM for the three models are plotted to be compared. The Fig. 5 shows the distributions associated with the evaluation of every sample for each digital staining model, through boxplots superimposed on violin graphics, individual points and outliers.

In the plot, the box covers the interquartile range (IQR) with the whiskers located at “minimum” ($Q1 - 1.5 \times \text{IQR}$) and “maximum” ($Q3 + 1.5 \times \text{IQR}$). The outliers are located farther than the whiskers limits.

The graph in Fig. 5 shows the superiority (almost 30%) of the generic model based on cycleGAN, compared to pix2pix and CUT. In addition, it shows that the distribution of values is quite homogeneous and there are no underlying distributions related to the type of tissue used in the test dataset. On the contrary, the values obtained by the pix2pix model seem to corroborate the weakness

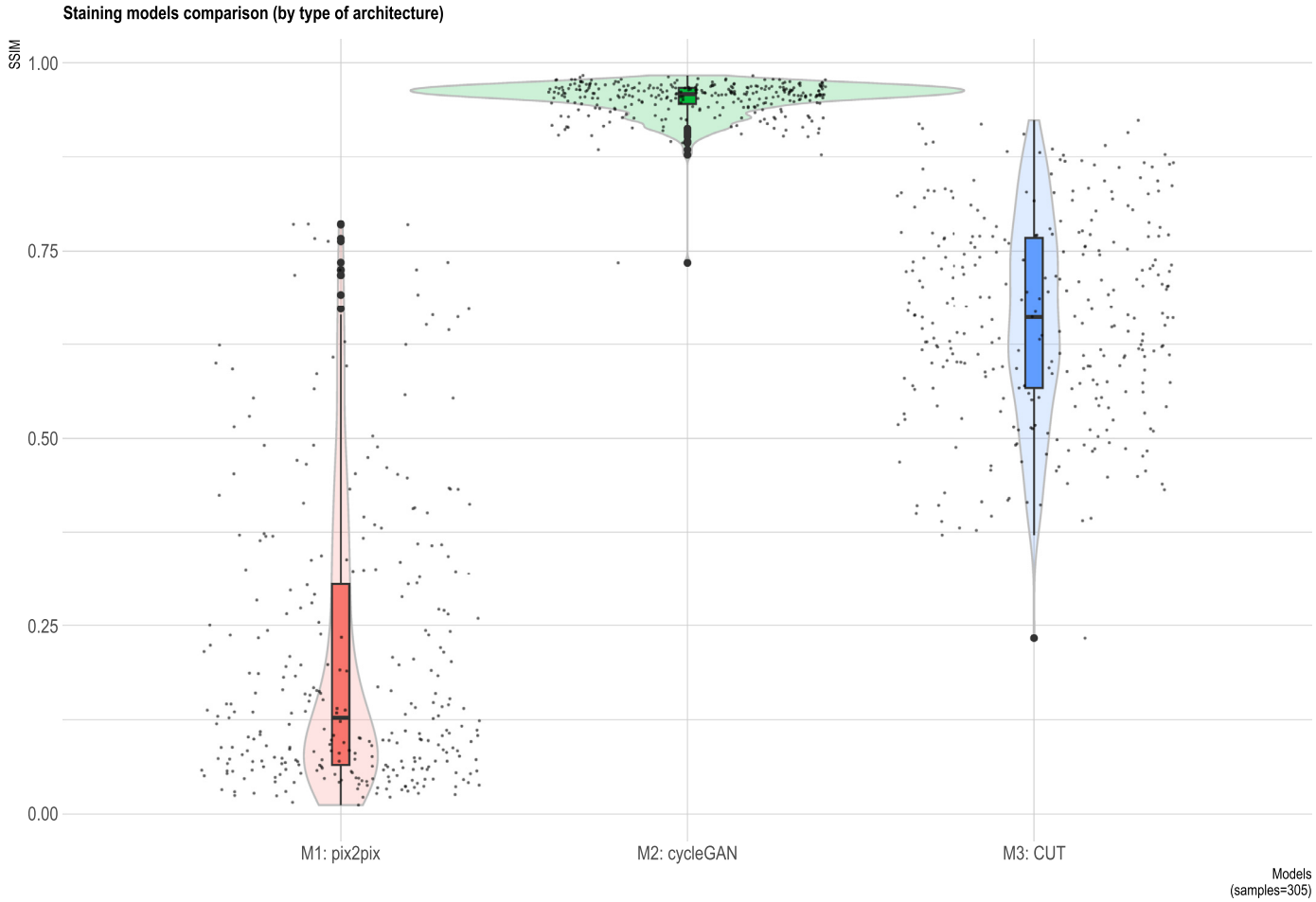

Fig. 5. Distribution plot comparison for structural similarity with different digital staining generic models. The confidence intervals (95%) for the means are: pix2pix (0.20559 ± 0.02147), cycleGAN (0.95261 ± 0.0026), and CUT (0.66012 ± 0.01525).

of this type of model when it is trained with images that are not perfectly aligned, as is the case with tissue samples.

3.2. (q2 answer) - Chromatic discrepancy

To compare the chromatic discrepancy obtained by the digital staining models with respect to the chemical stain used as a reference, it was decided to use a metric based on color distance. Instead of applying this metric at the pixel level, color quantization was used by clustering the color information in each image [51]. The elimination of the background pixels was accomplished before the clustering process was carried out in order to avoid their influence, since they do not correspond to stained tissue. In this way, only the relevant pixels are taken into account in the quantization process.

The clusters obtained in each pair of images (i.e., reference with chemical staining and correspondent to digital staining) were ordered by Hungarian algorithm so that the similar colors are closer. Finally, the distance between the clusters of both images is calculated using the EMD (Earth Mover's Distance metric, also known as Wasserstein distance) [52]. This metric not only takes into account the separation of the clusters in color space, but also their relative weight. In other words, this metric considers both the similarity of the colors in the images compared and their importance in the image as a whole.

When calculating the chromatic discrepancy between two images, it is worth studying the influence of two decisions that can affect the result, namely: the color space and the number of clusters. In reality, the choice of color space (i.e., RGB, HSV and CIELab) has no relevant influence, since more than absolute differences, we are interested in the relative differences in color obtained by the different staining models. This relevance has been tested empirically by comparing the distances obtained in different color spaces. Finally, after establishing the low influence of the color space chosen to compare the chromatic discrepancies achieved with the different models of digital staining, they were evaluated in the RGB color space.

To establish the number of clusters to quantize the color in the images, a study was conducted to compare the distances of the staining model based on cycleGAN against the reference chemical staining used in different number of clusters. The plot in Fig. 6 displays the distributions of the color discrepancy obtained based on the chosen number of clusters.

As illustrated in Fig. 3, staining variability in distinct regions of whole slide images (WSI) is intrinsically linked to tissue characteristics and their interaction with the staining process. Additionally, extrinsic factors arising from the staining process itself contribute to this variability. In order to address both types of variability, color normalization was applied prior to model training via experiments. For each tissue type, the normalization method proposed by Reinhard et al. [8] was chosen. However, to compare the chromatic discrepancy distributions obtained by digital staining models to those of the reference chemical staining, prior color normalization was disregarded.

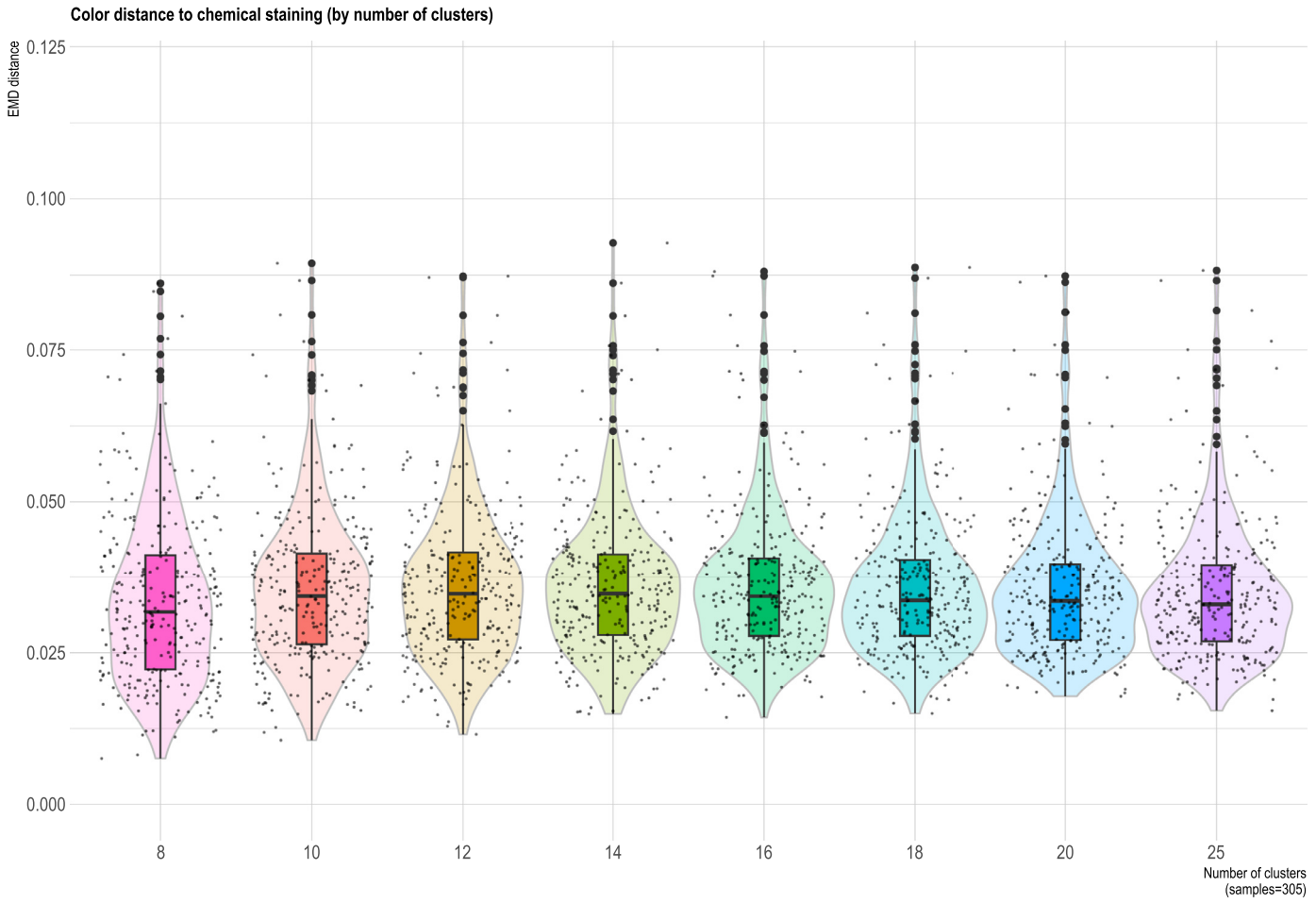

Fig. 6. Distribution plot comparison for color distance by number of clusters for digital staining generic cycleGAN based model.

In view of Fig. 6, we see that the increase in the number of clusters has a very limited effect except to reduce the dispersion of the distributions and increase the computation time.

Next, the chromatic discrepancy obtained with the three generic digital stain models was calculated to analyze their distribution in the set of test images. A number of 16 clusters were used for this calculation. The graphic in Fig. 7 shows how the best value around 10% (i.e., less chromatic discrepancy) is obtained using the cycleGAN based model. This should not be surprising since it is a model trained to exhibit cyclic consistency used for digital unstaining of samples. This graphic also shows that the other two models (i.e., pix2pix and CUT) have similar chromatic discrepancies, less than 20%, with respect to chemical staining.

3.3. (q3 answer) - Influence of tissue-specific modelling

Fig. 8 displays the comparison for structural similarity reached by the two kind of cycleGAN based models, the generic for any kind of tissue and the specific five (by type of tissue). This result allows addressing the question 3 regarding structural similarity. In such results, no significant improvements can be observed, referring to structural similarity, when digital staining models specific to the type of tissue are used.

Finally, the study was completed with the comparison of specific cycleGAN based models for each type of tissue to determine the possible influence of tissue type on the digital staining models. Fig. 9 shows the results obtained, in which it can be seen that the differences obtained are around 2% but are not very significant because overlapping is observed for the distributions we got.

3.4. (q4 answer) - Quality evaluation through subjective psychophysical tests

We conducted a qualitative evaluation using subjective quality metrics for the best model (cycleGAN) among the compared three quantitatively. This qualitative evaluation aims to measure the diagnosis capability of the digitally stained samples and the visual perceptual similarity between chemically and digitally stained samples. Thus, two experiments were performed: 1) Evaluation of 45 image pairs for each tissue type. The pair consists of a chemically stained sample and its digitally stained counterpart. So that, a total of 225 image sample were used. 2) Blind identification of the stain type of 64 sample images, (chemical or digital). Three experts were subjected to psychophysical tests for this purpose. We adhered to the ITU-R recommendation [53], whose main guidelines are: (1) The experts are subjected to an automated test consisting of a random series of 44 identical image pairs. (2) The pairs are displayed twice sequentially after a 10-second interval. (3) At the conclusion of each pair's display, there is a 10-second period devoted to assigning a score from 1 to 5 indicating how similar the observed pair of images are. In accordance with ITU-R, we referred to this metric as the Mean Opinion Score (MOS). (4) For the second experiment, the order of the images, chemical or digital stained image, was randomized to avoid any possible bias. (5) An additional case for each experiment is used as an example.

The rating scale utilized for the MOS is the absolute category rating, which maps ratings between 1 and 5 to the similarity of bad, poor, fair, good, and excellent, respectively. Also, according to

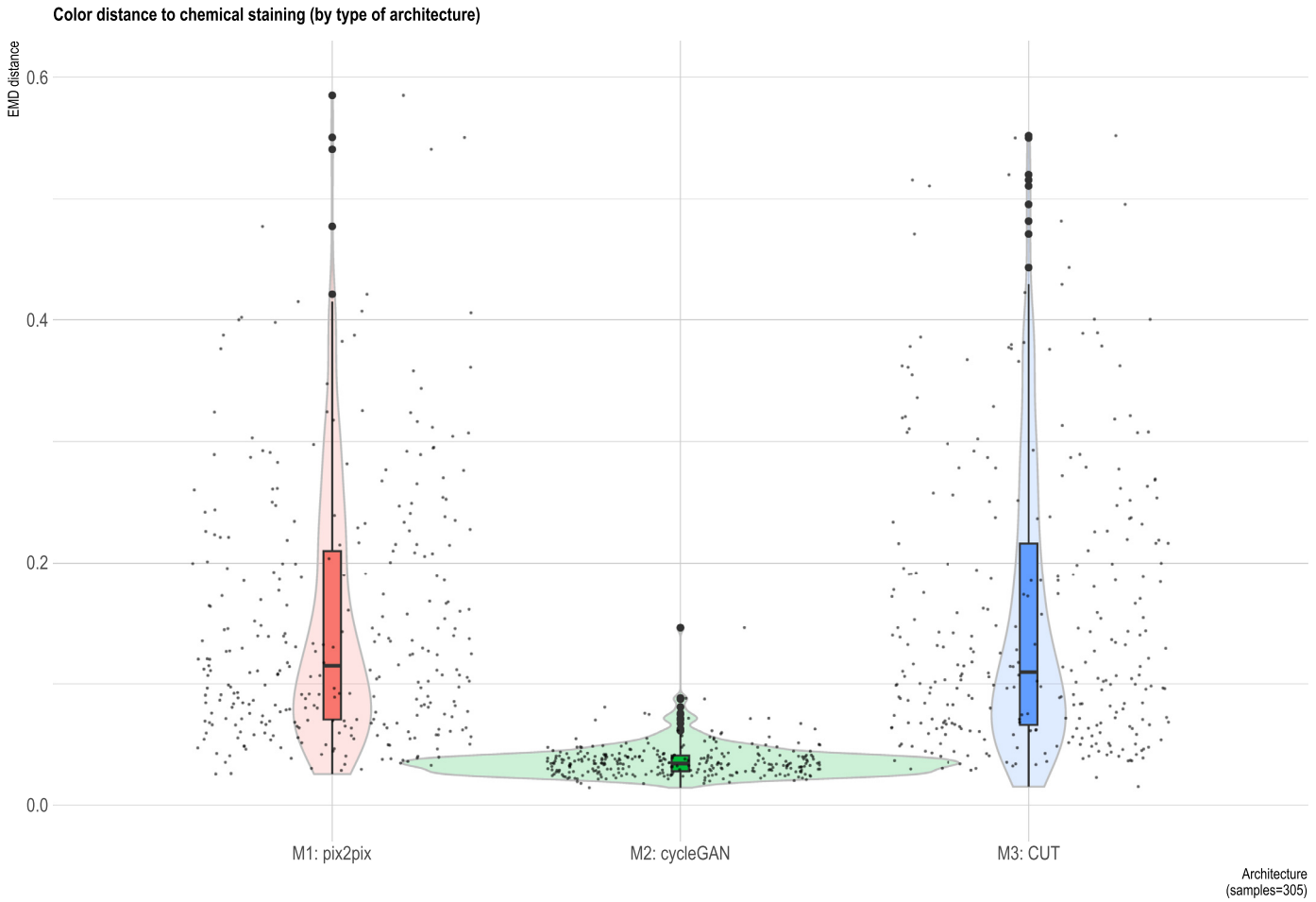

Fig. 7. Distributions plots comparison for color distance to reference chemical staining with different digital staining generic models. The confidence intervals (95%) for the means are: pix2pix (0.15017 ± 0.01179), cycleGAN (0.0364 ± 0.0015), and CUT (0.15628 ± 0.01368).

the experts, diagnostics are possible for images with a score of 4 or higher, despite the fact that images with scores of 4 may have color or contrast variations.

3.4.1. 1st experiment

It was conducted with 225 pairs of chemical and digital stained images. In the case of chemical stained samples 3.77% were scored with 1 and all corresponding to tumoral stroma tissue, 4.22% were scored with 2 where 3.12% were due to the score of the tumoral stroma tissue samples and 1.20% to the adipose tissue samples, 1.11% of samples were scored with 3 and all corresponding to the adipose tissue samples; 16% were scored with 4 and 74.88% were scored with 5, with similar percentages for all issue types. In the case of digital stained samples 5.55% were scored with 1, 2.66% were scored with 2, 4.88% of samples were scored with 3, 28.44% were scored with 4 and 58.44% were scored with 5. Observations revealed that the variation in scores was attributable to the same tissue types as those in chemically stained samples.

The perception of the quality of digitally stained images appears to be slightly inferior to that of chemically stained images. A Student's *t*-test was performed on the subset with scores 4 and 5, to determine if this difference is statistically significant and if it is possible to make a diagnosis independently of the image with digital or chemical staining. The means for these subsets of chemically and digitally stained samples were 45.44% and 43.44%, respectively and the test revealed that no significant difference exists.

Furthermore, a correlation between the MOS psychophysical scoring and the objective metric CPBD (Cumulative Perceptual Blurring Detection) ([53]) was calculated. The CPBD average value for the chemical stained samples was 0.6097, with the tumoral stroma tissue samples having the lowest value of 0.4481. The average CPBD value for digital stained samples was 0.5385, with the lowest value of 0.3618 due to tumoral stroma tissue samples. This is consistent with the MOS scoring. The Pearson linear correlation coefficient for chemically stained and digitally stained samples was 0.8941 and 0.8724, respectively. The calculated spearman rank-order correlation coefficients for both sets of stained samples were 0.8874 and 0.8733, with a mean absolute error of 0.1349 and 0.1761, respectively. As a result, the MOS and CPBD tests provide the same information and are consistent.

3.4.2. 2nd experiment

It was performed with 64 distinct images 32 chemically stained and 32 digitally stained samples. In this case 53.12% was true positive detections, that is the expert was able to distinguish properly the chemical stained samples, 34.37% were true negative detections that is the expert was able to distinguish properly the digital stained samples, 31.25% were false positives, that is they were identified as digital stained samples but they were chemical stained samples and 43.75% were false negatives, that is they were identified as chemical stained samples but they were digital stained samples. There was an average of 18.75% of samples were the expert could not make a decision whether they were chemical or digital stained samples.

These results were analyzed using the z-test to determine if there are statistically significant differences between the percep-

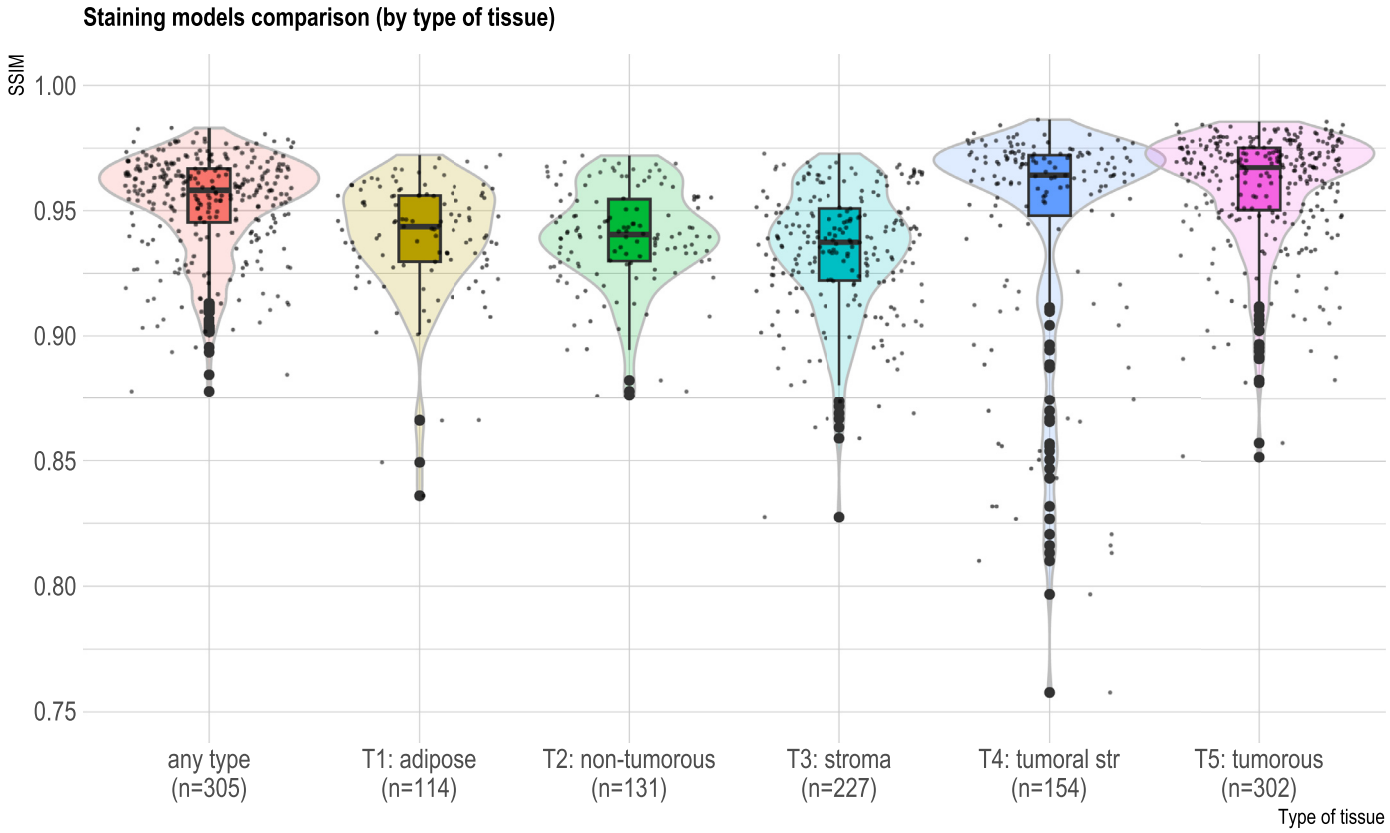

Fig. 8. Distribution plot comparison for structural similarity with digitial staining specific cycleGAN based models by type of tissue.

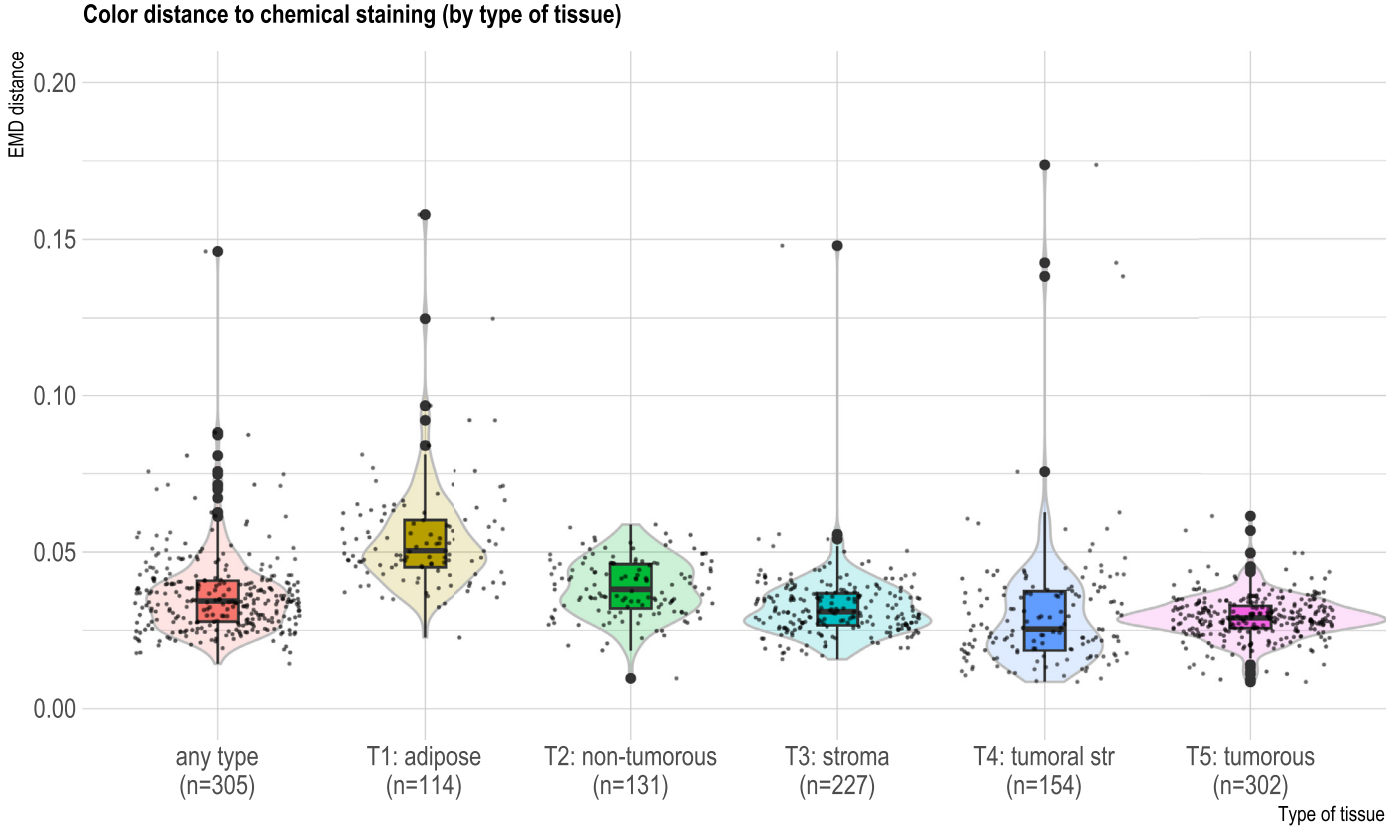

Fig. 9. Distributions plots comparison of color distance for digital staining specific cycleGAN based models (by type of tissue).

tion of chemically stained and digitally stained samples. The hypothesis that there are no significant differences between the perception of chemically stained images and digital images was accepted with a 95% level of confidence after tests were conducted considering both true and false detections. The z statistic was always less than 1.96, with a maximum value of 1.51.

3.5. Potential limitations

This study uses 13 pairs of samples (i.e., WSI) from which the patches that constitute the training dataset for the three models of digital staining are obtained (see Fig. 3). The final size of the training dataset is moderate but sufficient to obtain models with acceptable results (see Fig. 4). For the evaluation of these models, a test dataset was constructed from the ROIs of a WSI unseen by the models during training and even belonging to a different section of the tissues used previously. This test dataset consists of $1\,024 \times 1\,024$ px and 512×512 px patches, which are re-scaled to 256x256 px to meet the memory constraints of the available hardware. The proposed structural similarity and color discrepancy metrics evaluate the models with respect to the chemically stained images (i.e., ground truth) obtained on the re-scaled patches to provide a comparison unaffected by the aforementioned re-scaling. Furthermore, the quality evaluation of the obtained results from the best model did not show a significant degradation.

The models obtained have only been tested on breast tissues as they are the target of massive screenings in which automatic tools are of interest to reduce the analysis and decision-making time. However, the methodology followed in the presented study is extendable to other types of tissues. Additionally, the models can be adjusted for other types of tissues starting from the trained models by fine-tuning (i.e., transfer learning).

4. Conclusions

This paper has presented a comparison of three generative models of digital staining in H&E modality of breast tissue samples without staining, whose capture has been performed by multispectral microscopy reduced to three channels in the RGB visible spectrum. For the training of the models, samples with H&E chemical staining of the same type of tissues have been used, previously labeled —by a pathologist— in 5 types of tissue.

For training, one of the models (pix2pix) needs stained and unstained aligned images (i.e., paired), suffering from the inherent difficulty of obtaining aligned tissue samples. In this case, to achieve the necessary alignment of the samples, a computational process of the ROIs registration based on affine transformations was used. The other two models (cycleGAN and CUT) do not require aligned samples, so it was not necessary to use digital image registration.

The quantitative comparison of the three digital staining models was achieved in two dimensions: (1) structural similarity and (2) chromatic discrepancy. Both dimensions quantify the deviation that occurs in digital stain images from the reference represented by chemical stained images. In addition to the generic staining models (i.e., without distinction of tissue type), for cycleGAN and CUT, five specific models were generated for each type of tissue available with the intention of evaluating the effect of tissue specialization.

Alignment between reference (i.e., chemically stained) and perturbed (i.e., digital stain) images is required, since the measure of structural similarity (i.e., SSIM) and chromatic discrepancy (i.e., color distance) are considered as measures of deviation from a reference image. To obtain such alignment, the reference images have been subjected to a previous unstaining digital process applying the model that guarantees the cyclic consistency of the generative model based on cycleGAN. These images without staining are those that are subsequently digitally stained with each of the models and to which the metrics of structural similarity and chromatic discrepancy are applied.

To obtain the chromatic discrepancy value, a color distance was used based on the evaluation of the EMD distance between the color clusters associated with the tissue in each pair of images (with chemical and digital staining), which not only considers the color differences but also their proportion in the sample.

The main differences between the models are observed in the measure of structural similarity (see Fig. 5) in which cycleGAN obtains the best SSIM values, slightly higher than 0.95 (mean value). On the contrary, the values obtained with the pix2pix model reflect the weakness of these models in situations where it is not possible to have perfectly aligned pairs of images for training.

Regarding the measurement of chromatic discrepancy, a better result (around 10%) of the model based on cycleGAN with a distribution that presents a lower dispersion is also observed (see Fig. 9). In short, it is the model that obtains a color distribution most consistent with chemical staining.

When analyzing the graphics of the Figs. 5 and 7, it is observed that the distributions of the results does not present multimodality, so it is not appreciated that there are underlying distributions related to the type of tissue. That is, generic models of staining are insensitive to the type of tissue to be stained. This hypothesis is confirmed in digital staining with cycleGAN based models specific to the 5 available tissue types. When checking the distributions of structural similarity (see Fig. 8) and color distance (see Fig. 9) it is observed that the distributions overlap around very close central values. For this reason, the application of specific models by type of tissue should be ruled out, since they do not provide substantial improvements over generic models that are able to “learn” the characteristics that differentiate the types of tissues.

The results obtained with digital staining models highlight the potential value of reproducing automatic classification outcomes based on chemical staining samples. However, it is important to clarify that the primary objective of our study was to utilize digital staining as a means of enhancing the interpretability of results obtained through automated diagnostic and classification models. If such models could rapidly generate results using label-free input images, it would be highly beneficial for them to also provide modalities of comprehensible information (e.g., H&E staining) for specialists to confirm diagnoses. Thus, our study aims to compare three generative models to evaluate which one most accurately replicates the H&E staining modality. We conducted subjective psychophysical tests with three experts to qualitatively evaluation of the digital staining results obtained with the best model among the three quantitatively compared in our study. This qualitative evaluation provides evidence supporting the usability of digital staining as a promising tool to complement the diagnosis process.

The possibility of using a digital unstaining model on samples previously stained by a chemical or digital process raises questions that deserve a more detailed analysis in future work on the subject:*Would it be possible to use digital unstaining samples for training staining models that require image alignment? Would it be possible to obtain digital unstaining models that provide a universal result regardless of the original staining domain used?*

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

This work was funded by the HYPERDEEP project (Ref. SBPLY/19/180501/000273) supported by the Autonomous Government of Castilla-La Mancha and the HANS project (Ref. PID2021-127567NB-I00) supported by the Spanish Ministry of Science, Innovation, and Universities. All authors approved the version of the manuscript to be published. They would also like to extend the acknowledgment to Elena Ruiz y Alberto Velasco for their help in running some experiments.

References

[1] J.D. Bancroft, M. Gamble, Theory and practice of histological techniques, 8th edition, Elsevier health sciences, 2018.
[2] H.A. Alturkistani, F.M. Tashkandi, Z.M. Mohammedsaleh, Histological stains: a literature review and case study, Glob J Health Sci 8 (3) (2015) 72, doi:10.5539/gjhs.v8n3p72.
[3] Y. Rivenson, K. de Haan, W.D. Wallace, A. Ozcan, Emerging advances to transform histopathology using virtual staining, BME Frontiers 2020 (2020) 1–11, doi:10.34133/2020/9647163.
[4] Y. Zhang, K. de Haan, Y. Rivenson, J. Li, A. Delis, A. Ozcan, Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue, Light: Science & Applications 9 (1) (2020), doi:10.1038/s41377-020-0315-y.
[5] Z. Xu, X. Huang, C.F. Moro, B. Bozóky, Q. Zhang, GAN-based virtual re-staining: a promising solution for whole slide image analysis, arXiv (2019), doi:10.48550/ARXIV.1901.04059.
[6] K. de Haan, Y. Zhang, J.E. Zuckerman, T. Liu, A.E. Sisk, M.F.P. Diaz, K.-Y. Jen, A. Nobori, S. Liou, S. Zhang, R. Riahi, Y. Rivenson, W.D. Wallace, A. Ozcan, Deep learning-based transformation of H&Estained tissues into special stains, Nat Commun 12 (1) (2021), doi:10.1038/s41467-021-25221-2.
[7] S. Roy, A.k. Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, Micron 114 (2018) 42–61, doi:10.1016/j.micron.2018.07.005.
[8] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput Graph Appl 21 (4) (2001) 34–41, doi:10.1109/38.946629.
[9] 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, 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE, 2009, doi:10.1109/isbi.2009.5193250.
[10] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, IEEE Trans. Biomed. Eng. 61 (6) (2014) 1729–1738, doi:10.1109/tbme.2014.2303294.
[11] 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 Imaging 35 (8) (2016) 1962–1971, doi:10.1109/tmi.2016.2529665.
[12] B. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Holler, A. Homeyer, N. Karssemeijer, J. van der Laak, Stain specific standardization of whole-slide histopathological images, IEEE Trans Med Imaging 35 (2) (2016) 404–415, doi:10.1109/tmi.2015.2476509.
[13] A. Janowczyk, A. Basavanhally, A. Madabhushi, Stain normalization using sparse AutoEncoders (StaNoSA): application to digital pathology, Computerized Medical Imaging and Graphics 57 (2017) 50–61, doi:10.1016/j.compmedimag.2016.05.003.
[14] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization (2017), doi:10.48550/ARXIV.1703.06868.
[15] L.A. Gatys, A.S. Ecker, M. Bethge, A neural algorithm of artistic style (2015), doi:10.48550/ARXIV.1508.06576.
[16] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution (2016), doi:10.48550/ARXIV.1603.08155.
[17] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks (2014), doi:10.48550/ARXIV.1406.2661.
[18] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, doi:10.1109/cvpr.2017.632.
[19] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, doi:10.1109/iccv.2017.244.
[20] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks (2018), doi:10.48550/ARXIV.1812.04948.
[21] M.T. Shaban, C. Baur, N. Navab, S. Albarqouni, StainGAN: Stain style transfer for digital histological images (2018), doi:10.48550/ARXIV.1804.01601.
[22] T. Park, A.A. Efros, R. Zhang, J.Y. Zhu, Contrastive learning for unpaired image-to-image translation (2020), doi:10.48550/ARXIV.2007.15651.
[23] A. Bentaieb, G. Hamarneh, Adversarial stain transfer for histopathology image analysis, IEEE Trans Med Imaging 37 (3) (2018) 792–802, doi:10.1109/tmi.2017.2781228.
[24] M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: Overview, challenges and future (2017), doi:10.48550/ARXIV.1704.06825.
[25] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med Image Anal 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014), doi:10.48550/ARXIV.1409.1556.
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.
[28] A.V.D. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding (2018), doi:10.48550/ARXIV.1807.03748.
[29] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, IEEE, 2003, doi:10.1109/acssc.2003.1292216.
[30] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, doi:10.1109/tip.2003.819861.
[31] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444, doi:10.1109/TIP.2005.859378.
[32] L. Zhang, L. Zhang, X. Mou, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386, doi:10.1109/tip.2011.2109730.
[33] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 586–595, doi:10.1109/CVPR.2018.00068.
[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., Red Hook, NY, USA, 2012, pp. 1097–1105, doi:10.5555/2999134.2999257.
[35] J.J. Levy, C.R. Jackson, A. Sriharan, B.C. Christensen, L.J. Vaickus, Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized NCI Cancer Center, bioRxiv (2020), doi:10.1101/2020.01.07.897801.
[36] D. Li, H. Hui, Y. Zhang, W. Tong, F. Tian, X. Yang, J. Liu, Y. Chen, J. Tian, Deep learning for virtual histological staining of bright-field microscopic images of unlabeled carotid artery tissue, Molecular Imaging and Biology 22 (5) (2020) 1301–1309, doi:10.1007/s11307-020-01508-6.
[37] M. Lucic, K. Kurach, M. Michalski, S. Gelly, O. Bousquet, Are gans created equal? a large-scale study (2017), doi:10.48550/ARXIV.1711.10337.
[38] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C.H. van de Kaa, P. Bult, B. van Ginneken, J. van der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci Rep 6 (1) (2016) 26286, doi:10.1038/srep26286.
[39] P.-H.C. Chen, K. Gadepalli, R. MacDonald, Y. Liu, S. Kadowaki, K. Nagpal, T. Kohlberger, J. Dean, G.S. Corrado, J.D. Hipp, C.H. Mermel, M.C. Stumpe, An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis, Nat. Med. 25 (9) (2019) 1453–1457, doi:10.1038/s41591-019-0539-7.
[40] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: a survey, registration, IEEE Trans Med Imaging 32 (7) (2013) 1153–1190, doi:10.1109/TMI.2013.2265603.
[41] C.-W. Wang, S.-M. Ka, A. Chen, Robust image registration of biological microscopic images, Sci Rep 4 (1) (2014), doi:10.1038/srep06050.
[42] G. Haskins, U. Kruger, P. Yan, Mach Vis Appl 31 (1–2) (2020), doi:10.1007/s00138-020-01060-x.
[43] Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydın, J.E. Zuckerman, T. Chong, A.E. Sisk, L.M. Westbrook, W.D. Wallace, A. Ozcan, Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning, Nat. Biomed. Eng. 3 (6) (2019) 466–477, doi:10.1038/s41551-019-0362-y.
[44] X. Yang, B. Bai, Y. Zhang, Y. Li, K. de Haan, T. Liu, A. Ozcan, Virtual stain transfer in histology via cascaded deep neural networks, ACS Photonics 9 (9) (2022) 3134–3143, doi:10.1021/acsphotonics.2c00932.
[45] M. Boktor, B.R. Ecclestone, V. Pekar, D. Dinakaran, J.R. Mackey, P. Fieguth, P.H. Reza, Virtual histological staining of label-free total absorption photoacoustic remote sensing (TA-PARS), Sci Rep 12 (1) (2022) 10296, doi:10.1038/s41598-022-14042-y.
[46] B. Bai, H. Wang, Y. Li, K. de Haan, F. Colonnese, Y. Wan, J. Zuo, N.B. Doan, X. Zhang, Y. Zhang, J. Li, X. Yang, W. Dong, M.A. Darrow, E. Kamangar, H.S. Lee, Y. Rivenson, A. Ozcan, Label-free virtual HER2 immunohistochemical staining of breast tissue using deep learning, BME Frontiers 2022 (2022), doi:10.34133/2022/9786242.
[47] A. Rana, A. Lowe, M. Lithgow, K. Horback, T. Janovitz, A.D. Silva, H. Tsai, V. Shanmugam, A. Bayat, P. Shah, Use of Deep Learning to develop and analyze computational Hematoxylin and Eosin staining of prostate core biopsy

images for tumor diagnosis, JAMA Network Open 3 (5) (2020), doi:10.1001/jamanetworkopen.2020.5111. E205111
[48] S. Liu, B. Zhang, Y. Liu, A. Han, H. Shi, T. Guan, Y. He, Unpaired stain transfer using pathology-consistent constrained generative adversarial networks, IEEE Trans Med Imaging 40 (8) (2021) 1977–1989, doi:10.1109/tmi.2021.3069874.
[49] X. Li, H. Liu, X. Song, B.C. Brott, S.H. Litovsky, Y. Gan, Structural constrained virtual histology staining for human coronary imaging using deep learning (2022), doi:10.48550/ARXIV.2211.06737.
[50] N. Bayramoglu, M. Kaakinen, L. Eklund, J. Heikkila, Towards virtual H&E staining of hyperspectral Lung histology images using conditional generative adversarial networks, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), IEEE, 2017, doi:10.1109/iccvw.2017.15.
[51] M.E. Celebi, Improving the performance of k-means for color quantization, Image Vis Comput 29 (4) (2011) 260–271, doi:10.1016/j.imavis.2010.10.002.
[52] Y. Rubner, C. Tomasi, L.J. Guibas, The Earth Mover's Distance as a metric for image retrieval, Int J Comput Vis 40 (2) (2000) 99–121, doi:10.1023/A:1026543900054.
[53] R. Redondo, G. Bueno, G. Cristóbal, J. Vidal, O. Déniz, M. García-Rojo, C. Murillo, F. Relea, J. González, Quality evaluation of microscopy and scanned histological images for diagnostic purposes, Micron 43 (2) (2012) 334–343, doi:10.1016/j.micron.2011.09.010.
[54] A. Lahiani, J. Gildenblat, I. Klaman, S. Albarqouni, N. Navab, E. Klaiman, Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach, in: Digital Pathology, Springer International Publishing, 2019, pp. 47–55, doi:10.1007/978-3-030-23937-4_6.