



Staining Invariant Features for Improving Generalization of Deep Convolutional Neural Networks in Computational Pathology

Sebastian Otálora^{1,2*}, Manfredo Atzori¹, Vincent Andrearczyk¹, Amjad Khan^{1,3} and Henning Müller^{1,4}

¹ Institute of Information Systems, HES-SO University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland, ² Computer Science Centre (CUI), University of Geneva, Geneva, Switzerland, ³ Institute of Pathology, University of Bern, Bern, Switzerland, ⁴ Medical Faculty, University of Geneva, Geneva, Switzerland

One of the main obstacles for the implementation of deep convolutional neural networks (DCNNs) in the clinical pathology workflow is their low capability to overcome variability in slide preparation and scanner configuration, that leads to changes in tissue appearance. Some of these variations may not be included in the training data, which means that the models have a risk to not generalize well. Addressing such variations and evaluating them in reproducible scenarios allows understanding of when the models generalize better, which is crucial for performance improvements and better DCNN models. Staining normalization techniques (often based on color deconvolution and deep learning) and color augmentation approaches have shown improvements in the generalization of the classification tasks for several tissue types. Domain-invariant training of DCNN's is also a promising technique to address the problem of training a single model for different domains, since it includes the source domain information to guide the training toward domain-invariant features, achieving state-of-the-art results in classification tasks. In this article, deep domain adaptation in convolutional networks (DANN) is applied to computational pathology and compared with widely used staining normalization and color augmentation methods in two challenging classification tasks. The classification tasks rely on two openly accessible datasets, targeting Gleason grading in prostate cancer, and mitosis classification in breast tissue. The benchmark of the different techniques and their combination in two DCNN architectures allows us to assess the generalization abilities and advantages of each method in the considered classification tasks. The code for reproducing our experiments and preprocessing the data is publicly available¹. Quantitative and qualitative results show that the use of DANN helps model generalization to external datasets. The combination of several techniques to manage color heterogeneity suggests that several methods together, such as color augmentation methods with DANN training, can generalize even further. The results do not show a single best technique among the considered methods, even when combining them. However, color augmentation and DANN training obtain most often the best results (alone or combined with color normalization and color augmentation). The statistical significance of the results and the embeddings visualizations provide useful insights to design DCNN

OPEN ACCESS

Edited by:

Anne Martel,
Sunnybrook Research Institute (SRI),
Canada

Reviewed by:

Metin Gurcan,
Wake Forest University, United States
Pengyue Zhang,
Stony Brook University, United States

*Correspondence:

Sebastian Otálora
Juan.Otalora@etu.unige.ch

Specialty section:

This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Bioengineering and
Biotechnology

Received: 27 May 2019

Accepted: 05 August 2019

Published: 23 August 2019

Citation:

Otálora S, Atzori M, Andrearczyk V,
Khan A and Müller H (2019) Staining
Invariant Features for Improving
Generalization of Deep Convolutional
Neural Networks in Computational
Pathology.
Front. Bioeng. Biotechnol. 7:198.
doi: 10.3389/fbioe.2019.00198

¹https://github.com/sebastianffx/stain_adversarial_learning

that generalizes to unseen staining appearances. Furthermore, in this work, we release for the first time code for DANN evaluation in open access datasets for computational pathology. This work opens the possibility for further research on using DANN models together with techniques that can overcome the tissue preparation differences across datasets to tackle limited generalization.

Keywords: staining normalization, adversarial neural networks, digital pathology, color augmentation, color normalization, domain shift

1. INTRODUCTION

Since its start, one of the main goals of computational pathology (CP) is to find precise and reproducible methods to quantify the content of tissue slides and the relationships of this with the disease stage and patient outcome (Madabhushi, 2009; Madabhushi et al., 2011; Al-Janabi et al., 2012; Kothari et al., 2013). During the last decade the methods for analyzing images in digital pathology have become more precise and have greatly benefited the CP community, also thanks to the steady development of deep learning algorithms and particularly thanks to deep convolutional neural networks (DCNN), Shifting from handcrafted features toward end-to-end architectures [to detect cancer in histopathology images at the image patch level (Veta et al., 2015; Janowczyk and Madabhushi, 2016; Ciompi et al., 2017) and at the whole-slide-image level (Litjens et al., 2016; Cruz-Roa et al., 2018)], methods have become more precise, achieving in some cases for specific tasks classification performance comparable to pathologists.

Despite the performance improvements of the methods, there are still technical barriers that prevent the translation of these advances into better clinical applications. Two of the most typical chemicals used in pathology to stain tissue slides are Hematoxylin and Eosin. These chemicals highlight the nuclei with a dark purple color (Hematoxylin) and the cytoplasm with a light pink one (Eosin). One of the most important factors preventing the application of machine learning methods to clinical practice is related to the heterogeneity of Hematoxylin and Eosin (H&E) images due to tissue preparation and several parameters involved in the tissue preparation and digital scanning process (temperature of the tissue, thickness of the cuts, image sensor of the digital camera, etc.). Several image processing and machine learning techniques reported in the literature deal with color heterogeneity improving classification and segmentation performance for various tissue types (Van Eycke et al., 2017; Roy et al., 2018; Tellez et al., 2019). However, this challenging problem is far from being solved.

Color heterogeneity can affect the performance of the machine learning algorithms that easily overfit when trained with data from one center or scanner and fail to generalize to images from other centers (Kothari et al., 2014; Ciompi et al., 2017). One example of this problem is shown in **Figure 1**.

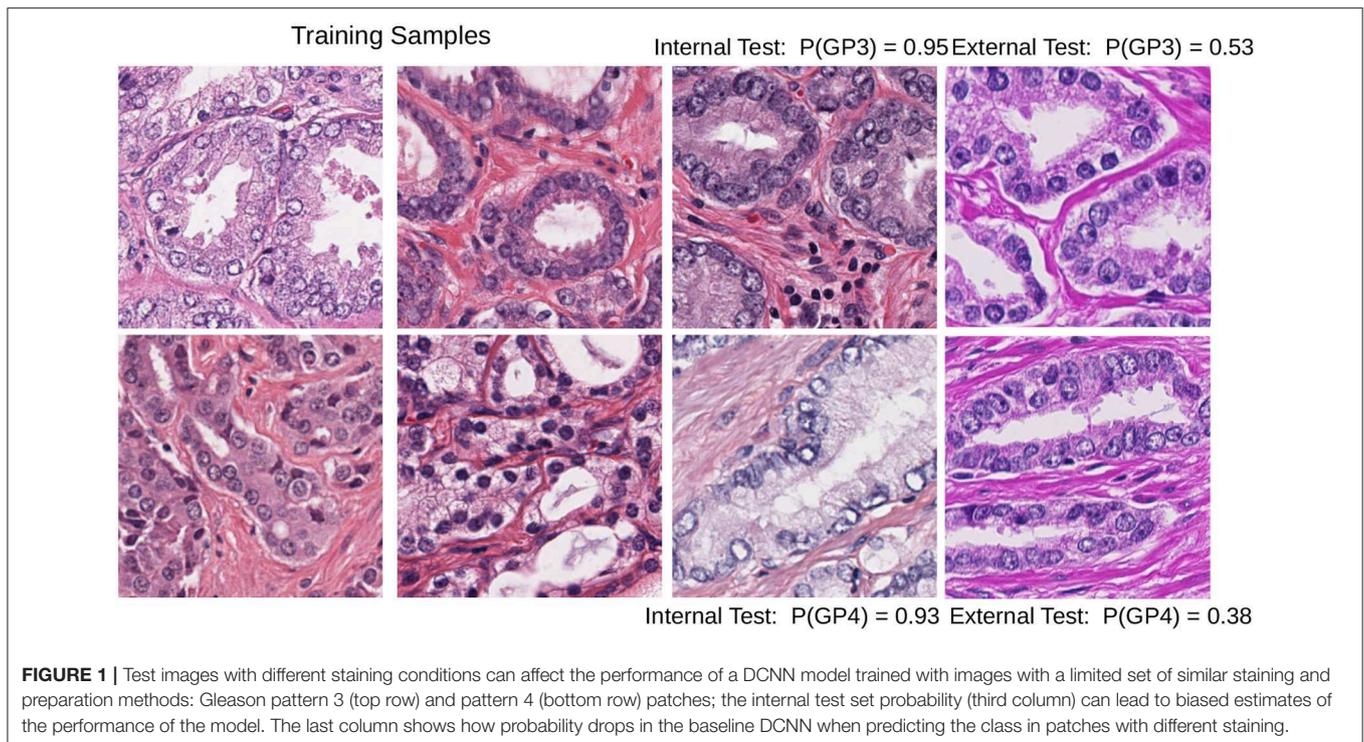
The heterogeneity of colors in H&E images is due to tissue preparation and is related to the complex set of preparation phases related to staining procedures, section thickness, and scanner differences (Leo et al., 2016).

One of the more recurrent problems in the preparation of the specimen is over- or under-staining. The time and amount of dye applied to the tissue on the glass make the process prone to perceptual differences in color and intensity of the tissue in the digitized slide (Bejnordi et al., 2016; Van Eycke et al., 2017).

The changes in color intensity due to different slice thicknesses are acknowledged as a source of variation. Few articles have explored how to overcome this aspect using automatic tools, presumably because of the complexity of the experimental setups. Notably, the work of Bug et al. (2017) provides a staining normalization method based on an end-to-end deep learning (DL) architecture that takes into account the context of the tissue to normalize the extracted features. The authors evaluated their method on lung cancer tissue images varying not only in terms of H&E concentrations but also in section thickness. They showed that their method gives a consistent normalization through the studied protocols and less variance in the output with respect to the classic methods of Macenko et al. (2009) and Bejnordi et al. (2016).

Although standardization procedures are often applied in laboratories and clinical practice, perfect color calibration among samples is hard to achieve. The scanning parameters that are hardcoded in the whole slide scanner hardware (that change according to the vendors, such as Aperio, Philips, Ventana) result in specific color characteristics in the resulting digital image, for example, varying in the image sensor and also the stitching techniques. These parameters are also a part of the digital pathology pipeline on which pathologists have limited control. In the work of Leo et al. (2016) the authors evaluate the stability of features in prostate cancer classification across several scanner producers. Their results show that only a portion of the commonly used features in digital pathology are robust to such scanner differences. Interestingly, the authors show that color normalization alone cannot solve the problem of inter-scanner feature instability. Therefore, it is crucial to develop algorithms that generalize well on heterogeneous datasets for a more robust deployment of computer-aided diagnostic systems (Kothari et al., 2013; Leo et al., 2016).

Image processing and machine learning approaches that deal with color heterogeneity show generalization improvements in classification and segmentation performance (McCann et al., 2014; Vahadane et al., 2016; Bentaieb and Hamarneh, 2018; Ren et al., 2018; Roy et al., 2018; Tellez et al., 2018a,b, 2019). Such techniques can generally be grouped into three different types. The first approach (which is probably the most frequently studied) is staining normalization, meaning that the color



concentration of all the images in the database is mapped to match the staining appearance of a target image. This problem has been thoroughly studied in literature (Macenko et al., 2009; Li et al., 2015; Ciompi et al., 2017; Van Eycke et al., 2017). Staining normalization applied with DCNN allowed to improve classification accuracy by over 20% (Ciompi et al., 2017) in colorectal cancer tissue classification. In the work of Van Eycke et al. (2017) a series of steps for color vector extraction is described. Two main challenges often arise when using staining normalization techniques. First, as the source images are color-transformed into the target image color space, the algorithms using the normalized images are sensitive to the selection of the image taken as reference (Bentaieb and Hamarneh, 2018). Second, the time of staining normalization algorithms for standardizing a batch of images can easily take up to 30 minutes as noted by Shaban et al. (2018).

The second approach is color augmentation, where many variations of the original image are created for training by varying intensity, brightness, contrast, RGB channel values and also altering them in other color spaces such as the Hue Saturation Value (HSV) space, or deconvolving the Hematoxylin-Eosin-DAB (HED) channels and slightly modifying them (Van Eycke et al., 2018; Bandi et al., 2019; Tellez et al., 2019). The color augmented images are usually generated at training time and given as input to the algorithm to cope with possible variations in the test set. Data augmentation is usually built into the pipeline of training deep learning models, since such models usually learn color invariances by processing a large amount of annotated samples. Color augmentation can lead to excellent results, such as the ones obtained by the leading teams of the

Camelyon17 challenge (Bandi et al., 2019). These approaches use deep learning architectures in conjunction with extensive color augmentation to force their networks to be robust to color variation.

The third and most recent family of techniques is inherently related to deep learning architectures. Deep learning models account for the complexity of learning the staining transformation in the test set by means of learning a cascade of non-linear transformations of the training input images. This learnt information is used to either normalize the image or to explicitly design a deep learning model architecture that penalizes or does not take into account the staining information (Bug et al., 2017; Janowczyk et al., 2017; Ren et al., 2018). This deep learning-based family of techniques is further discussed in section 2.4.

As discussed, staining variability and color augmentation can dramatically affect automatic image analysis algorithms and several techniques have been proposed to tackle the problem depending on the image analysis task. Perfect color calibration among samples is hard to achieve, despite standardization procedures being applied in clinical practice. In this paper, we contribute to tackle this challenge by comparing the three mentioned approaches in intra and inter-center classification tasks (targeting the classification of mitotic cells in breast cancer and Gleason pattern classification in prostate cancer). We thoroughly evaluate the adversarial neural network training approach first proposed by Lafarge et al. (2017) to learn domain invariant features, showing that the use of DANNs can help generalization to external datasets. The combination of techniques also suggests that by using basic color augmentations

in addition to techniques such as DANN training and staining normalization, the models can generalize even further.

2. MATERIALS AND METHODS

2.1. Color-Heterogeneous Datasets

For comparing the generalization capabilities of the approaches, we used two highly heterogeneous data sets that target classification tasks. The data sets account for the staining variability by including images with differing preparation parameters from several centers. Such parameters result in strong differences in the visual characteristics of the images.

Most frequently, for the evaluation of algorithms, subsets of images from the same center are used as a training set and then validation and testing is done with images of different patients but with similar image preparation characteristics from the same center.

In our evaluation, we extend the classical evaluation by separating internal test data: different patients with similar preparation parameters, from external test data: different patients and preparation characteristics. Such an evaluation provide us with a better approximation of the performance of the algorithms on images acquired from different centers, i.e., the generalizability of the networks to such changes.

The first dataset is the Tumor Proliferation Assessment Challenge (TUPAC), which is built to evaluate algorithmic performance for mitotic figure detection in breast cancer tissue (Veta et al., 2019). We refer to this dataset as TUPAC. The dataset contains 1,522 mitotic figures extracted from high power fields of 73 breast cancer cases from three pathology labs. This suits our purpose of evaluating the generalization abilities of algorithms across inter-center variability. The dataset partition is the same as in Lafarge et al. (2017). All our DCNN models in this dataset are trained and validated with eight (458 mitoses) and four cases (92 mitoses) from the first pathology lab. The remaining 12 cases (533 mitoses) from the first pathology lab are used as an internal test set. The 50 examples from the two other pathology labs (469 mitoses) are used as an external test set. To create a set of challenging negative samples, an initial CNN is trained with mitotic patches as positive patches and on random patch locations as negative ones that do not overlap with mitotic patches. Once the network is trained, only the false positives and negative patches with a high probability of being mitotic are taken as negative samples, this set of hard-negative mined samples are kept throughout all the experiments. Our hard-negative mining differs, and thus, the performance is not directly comparable with those reported by Lafarge et al. (2017). The cardinality of each partition is in **Table 1**, and a few example image locations are shown in **Figure 2** for TCGA-PRAD and example image patches in **Figure 3** for the TUPAC dataset, respectively.

The second dataset is comprised of images from prostate cancer tissue with Gleason patterns 3 and 4 (GP3, GP4). It contains image patches from diagnostic slides of the cancer genome atlas prostate adenocarcinoma dataset (TCGA-PRAD) and also from the manually annotated prostate tissue microarray images used in the study of Arvaniti et al. (2018), we refer to this dataset as TCGA-TMAZ. The TCGA-PRAD images were recorded in 16 tissue source centers² that were used for training, validation and internal test sets. Whereas, the TMA-Zürich where

TABLE 1 | Number of original patches for the TUPAC dataset, due to the high class imbalance, we use data augmentation only in the mitotic class by creating rotated and flipped patches.

Partition	Mitosis	Non-mitosis	Total	# of cases	Center
Train	458	3,842	4,300	8	1
Validation	92	1,196	1,288	4	1
Internal test	533	12,317	12,850	12	1
External test	469	505	974	50	2,3
Total	1,552	17,860	19,412	74	2,3

²https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

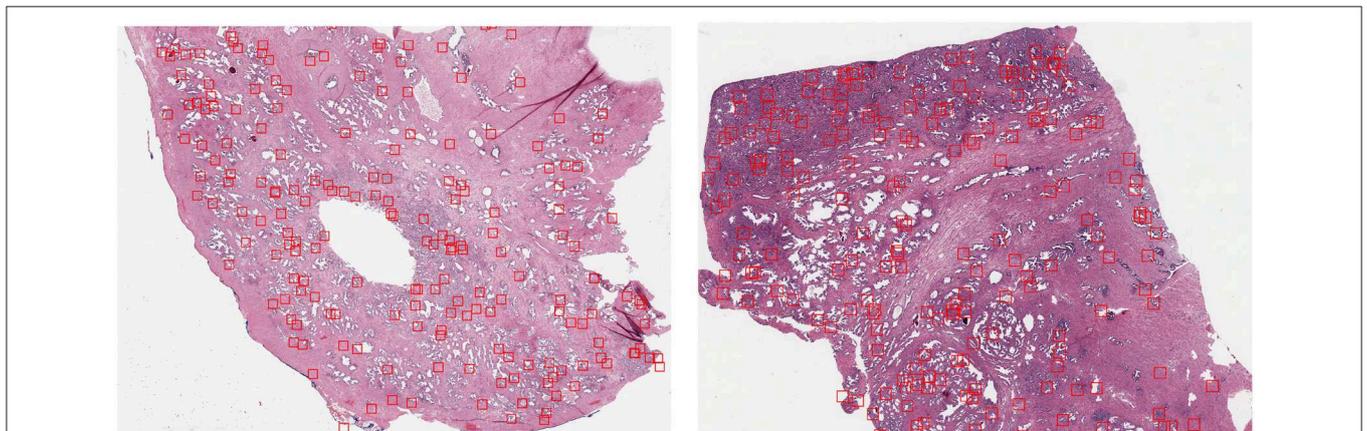
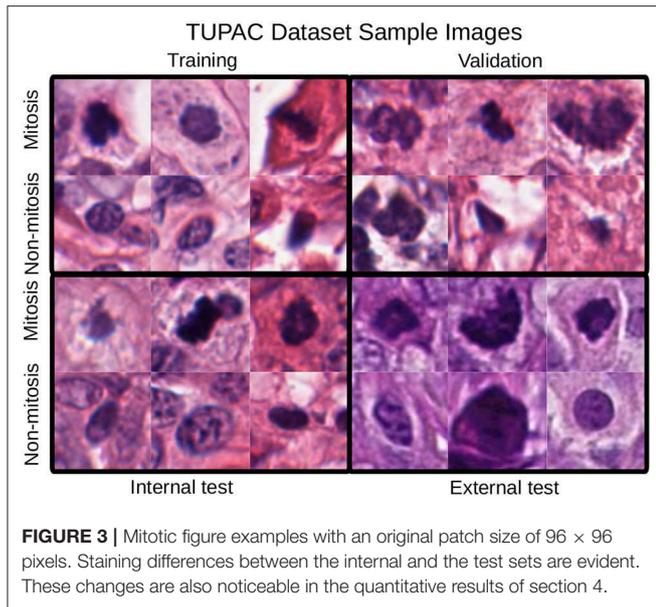


FIGURE 2 | GP3 patch locations extracted (red bounding boxes) from slide TCGA-2A-A8VL (Left) belonging to the training set and the slide TCGA-EJ-7321 (Right) from the internal test set using the heatmap resulting from Equation (2).



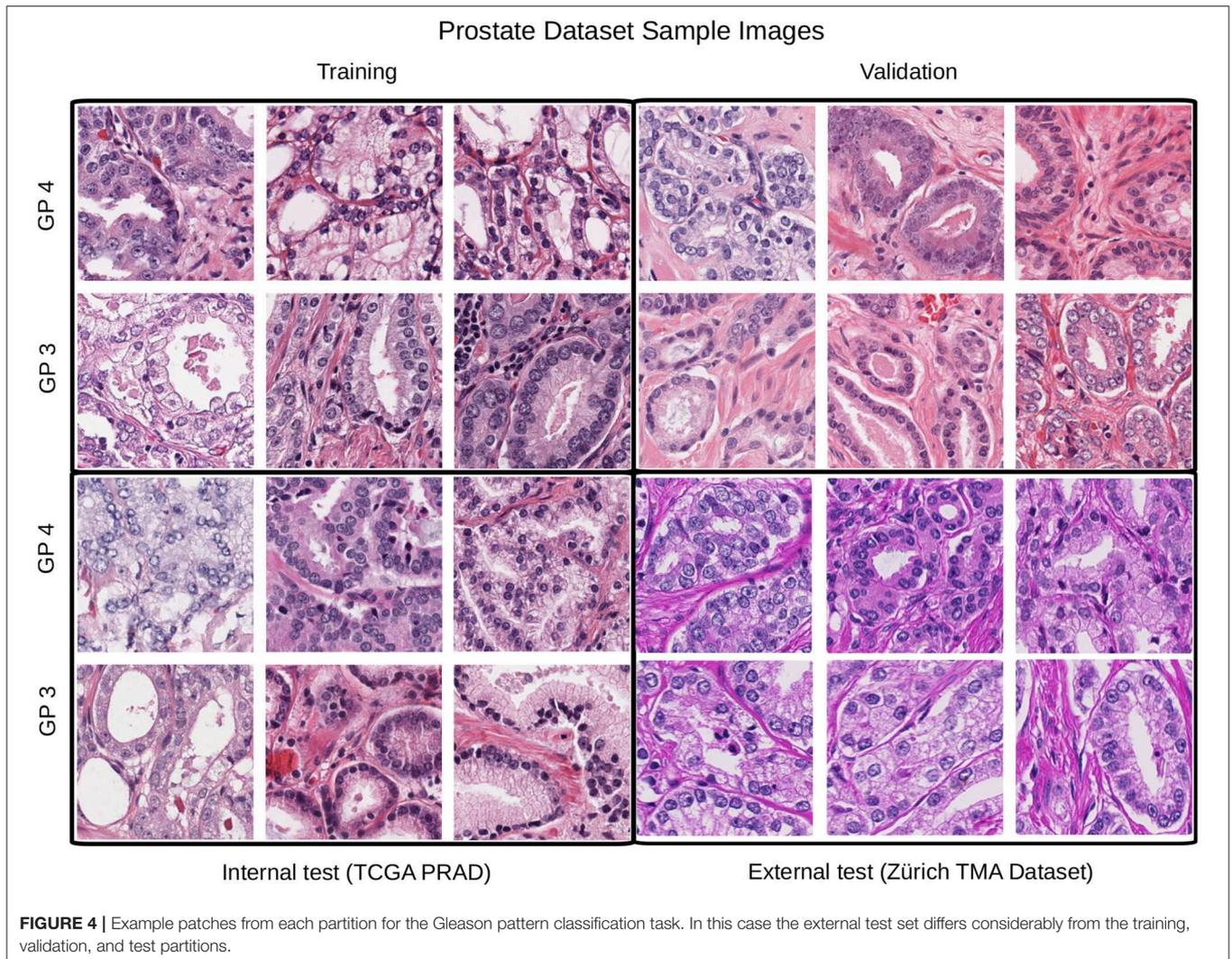
used as an external test set since it contains significant staining variability as shown in **Figure 4**.

Since in the TCGA-PRAD dataset there are no Gleason pattern annotations but only the global labels of the two most prominent Gleason patterns, we selected only those diagnostic images with the same primary and secondary Gleason patterns, i.e., GS = 3 + 3 and GS = 4 + 4. With this consideration, we are more likely to extract patches from relevant regions. Furthermore, we employ the heuristic first presented by Rousson et al. (2018), to build a heatmap that guides patch extraction from non-annotated Whole Slide Images (WSI) using the following transformation of the original RGB WSI at the 10X magnification level:

$$h_1 = \tanh\left(\frac{2(B - R)}{G + 1} + 0.5\right) + 0.5;$$

$$h_2 = \tanh\left(\frac{640 - R - G - B}{300} + 0.5\right) + 0.5 \quad (1)$$

$$h = 0.5 \tanh(h_1 h_2 - 1.75) + 0.5 \quad (2)$$



By fixing a threshold (0.65) on the resulting heatmap h we selected a fixed amount of random patches in the WSI from positive locations in the thresholded image. Example locations of extracted patches from two WSIs of the train and internal test sets are illustrated in **Figure 2**. While there are a minority of patch locations outside relevant regions, this unsupervised approach to locate regions of interest allows us to have the right amount of relevant regions that contains GP3 and GP4 patches.

In contrast to the TUPAC dataset, in this case, we have more than 3 centers, the total number of source centers for the TCGA-TMAZ dataset is 34, which makes it challenging to train the algorithms with a limited amount of images for each center. The total number of images per partition is reported in **Table 2**.

We do not have a larger public data set with information on the different centers, which is a current challenge for evaluation of the generalization of CP algorithms. In the literature there are interesting approaches to solve CP classification tasks that use more than 100 million annotated patches as in Nagpal et al. (2018). However, it is noteworthy that in this work we did not use any annotations of regions or private data source. Only the public annotations of the TUPAC mitotic regions and an open access dataset with limited annotations are used. The scripts and filename lists with (x,y) locations of the patches are provided in order to allow reproducing the same image patches used in our experiments.

2.2. Staining Normalization

Since all the pixels of the digital H&E images are represented in RGB space, ideally each pixel should contain a composition of the color representation of Hematoxylin, Eosin, and background. Images acquired from the same center and using the same preparation parameters should share the stain absorbance coefficients, which can be written as the linear transformation (omitting background that should be close to 255 for the three channels):

$$S = \begin{pmatrix} H_R & H_G & H_B \\ E_R & E_G & E_B \end{pmatrix}$$

Where the first-row vector corresponds to the RGB components of hematoxylin and the second one to the components of Eosin. In staining normalization methods, the aim is to estimate the individual staining absorbance coefficients of the images S and quantify the absorbed light C by the tissue when it was scanned, which is the value in the H&E space of each pixel. The Beer-Lambert law provides a way to estimate them in the optical density space, given the original pixel content for the c -channel I_c :

$$I_c = I_0 \exp(-S_c \cdot C)$$

Where c ranges in the RGB channels, $S \in [0, +\infty]^{3 \times 2}$ is the matrix of absorbance coefficients, $C \in [0, +\infty]^2$ is the vector of the two staining concentration coefficients and I_0 is the background value. The widely used method of Macenko et al. (2009) provides an estimation of S by computing a plane using the two largest singular value decomposition vectors of the image and then projecting the data into this plane and clipping extreme

TABLE 2 | Number of original patches for the TCGA-TMA dataset.

Partition	GP3	GP4	Total	#centers
Train	1,184	1,219	2,403	6
Validation	479	658	1,137	4
Internal test	811	510	1,321	6
External test	1,602	2,359	3,961	1
Total	4,076	4,746	8,822	17

values. Using this estimation, the matrix S is fixed to normalize all the images in the database by multiplying each value of the concentration of the source images pixelwise. Then, the image is in the H&E space of the target (template) image. An optional step is to preprocess the image with brightness standardization making the obtained coefficients less dependent on the brightness values. A schematic view of the normalization approach, used in our experiments, is displayed in **Figure 5**.

2.3. Color Augmentation

One common strategy used to create new images with color variations consists of multiplying each of the color channels I_c (or estimated staining concentrations in the H&E space) of the original image by a random small constant a_c that will scale the original intensity value and then add a second constant b_c that shifts the color toward higher/lower intensity values:

$$I'_c \leftarrow a_c I_c + b_c \tag{3}$$

The range from where the random constants a_c and b_c are drawn determines how much variation is allowed in the generation of the new images. Examples of color augmented patches from the TUPAC dataset using this strategy in the RGB space are shown in **Figure 6**.

While staining normalization methods aim at homogenizing the appearance of the images, they might fall short at inter-center generalization because of a non-optimal normalization or overfitting in the training set due to the small amount of variation between the samples. Recent studies have shown how by creating images for training with an expansion of the training set with duplicate images from a broader color range have a significant impact at generalization in the inter-center evaluation of machine learning methods. Particularly with deep learning architectures, studies have shown how color augmentation techniques usually outperform staining normalization methods (Bejnordi et al., 2017; Tellez et al., 2019). This might be intuitive in the context of training deep learning models, where the larger the amount of data the model is fed with, the more variations the model is exposed to. Therefore, it is more robust to changes in appearance in the test set.

2.4. Deep Learning Approaches for Staining Normalization

Recent approaches based on deep learning models such as staining normalization stacked autoencoders (Janowczyk et al., 2017) and U-net-based architectures (Tellez et al., 2019) allow to capture complex staining transformations and build normalizers and generators in an end-to-end manner. In the

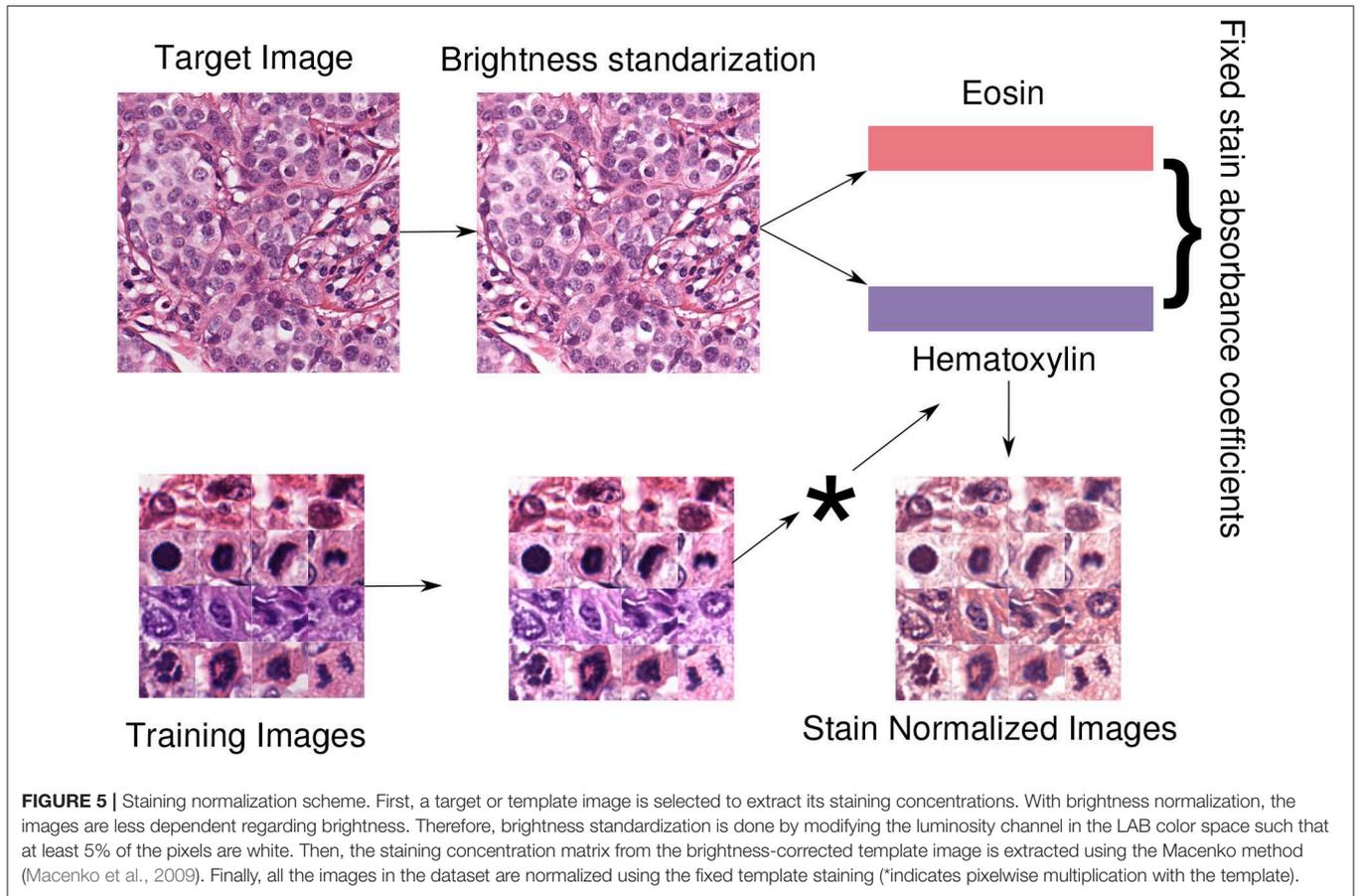


FIGURE 5 | Staining normalization scheme. First, a target or template image is selected to extract its staining concentrations. With brightness normalization, the images are less dependent regarding brightness. Therefore, brightness standardization is done by modifying the luminosity channel in the LAB color space such that at least 5% of the pixels are white. Then, the staining concentration matrix from the brightness-corrected template image is extracted using the Macenko method (Macenko et al., 2009). Finally, all the images in the dataset are normalized using the fixed template staining (*indicates pixelwise multiplication with the template).

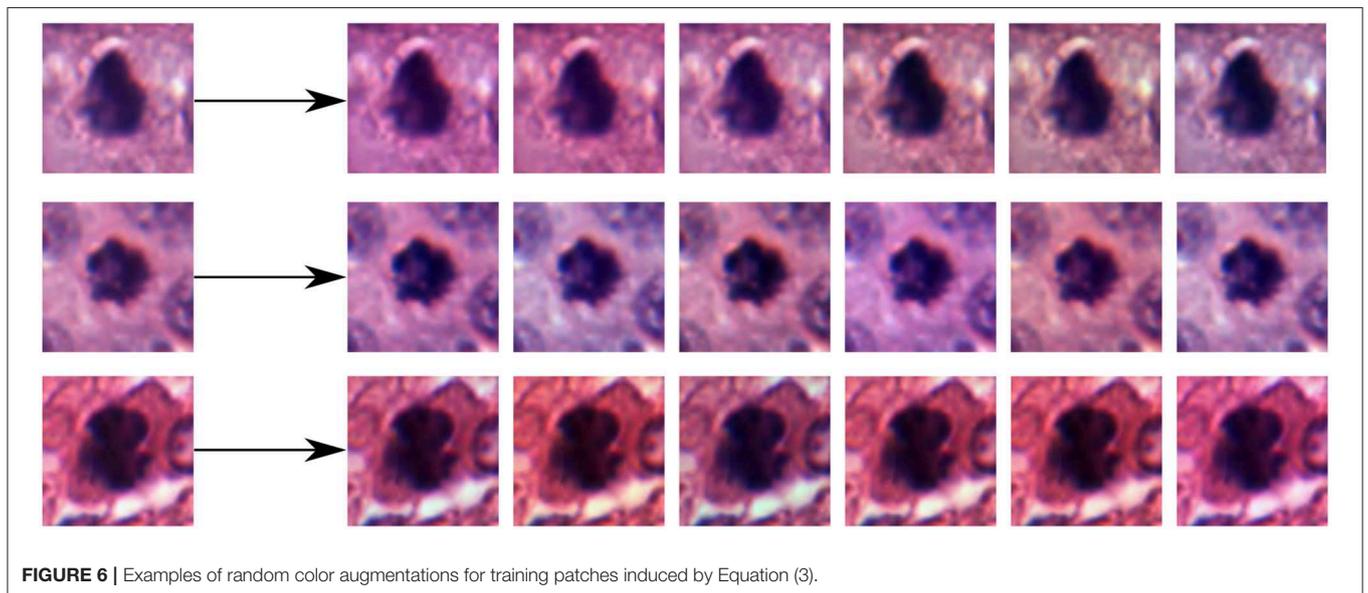


FIGURE 6 | Examples of random color augmentations for training patches induced by Equation (3).

work of Tellez et al. (2019) the impact of color augmentation and staining normalization is assessed in three organs regarding cancer: prostate, breast, and colorectal. The authors propose a normalization method based on a U-Net architecture that

mapped augmented versions of the image to a normalized one. This showed significantly better results than standard augmentation and normalization methods. In Bentaieb and Hamarneh (2018), the authors explored several ways of training

a deep adversarial staining transfer model in colon, ovary and breast cancer datasets showing that a combination of penalty terms of consistency of the normalized images and a conditional term that adapts to the task leads to the best results in classification tasks. The models mentioned above require the training of an external deep learning model that normalizes the images. This is used subsequently for the DCNN model training of the task, thus not using the source domain information for guiding the training of staining invariant features.

2.5. Domain Adaptation and Adversarial Learning

From a data-driven perspective, the disparities between the digital histopathology image stainings caused by the changes in preparation methods and scanners make the distribution of the generated images differ. The training of machine learning models with only images created from a subset of preparation methods, scanners or centers creates vulnerable models that might fail at correctly classifying near out-of-distribution samples with staining changes. Discriminative learning methods for classification, such as neural networks or support vector machines, perform well when training and test data are drawn from the same distribution. Therefore, when a model is trained using a set from one distribution and then tested in another, its performance will be hindered by the dissimilarity between the training and test distributions. This is a well-known problem in machine learning called domain adaptation (DA). It is an active research area in machine learning where many novel frameworks have been proposed (Crammer et al., 2006; Ben-David et al., 2007; Tzeng et al., 2017).

In Ben-David et al. (2007), the authors provided a framework to analyze the contributions of domain adaptation for the generalization of models by learning features that account for the domain disparity between training and test set distributions. The theory of DA suggests that a good representation for cross-domain transfer is one in which the algorithm cannot learn to identify the domain of origin of the input observation. This led to the authors of Ganin et al. (2016) to propose a concrete implementation of this idea in the context of deep neural network models. The objective of domain adversarial neural network models (DANN), is to learn features that do not take into account the domain of the training samples. The domain adversarial features combine discriminativeness and domain invariance into the same representation. To build such representations, two classifiers are involved: (i) the label classifier that predicts the main task classes (e.g., mitosis/non-mitosis) at training and testing time, and (ii) the domains classifier that discriminates between several domains. Both classifiers can be tied using the same set of features. Having two outputs with separate loss functions, one that measures the error at classifying the class of the sample correctly and the second one, measuring the error at classifying the origin of the sample (i.e., the domain).

The optimization objective of a DANN model is to find a saddle point solution of parameters for the task classifier $\hat{\theta}_y$, domain parameters $\hat{\theta}_d$ and domain-invariant features $\hat{\theta}_f$. The loss

function for the $N = n + n'$ training samples is:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \left(\frac{1}{n} \sum_{i=1}^n \mathcal{L}_d^i(\theta_f, \theta_d) + \frac{1}{n'} \sum_{i=n+1}^N \mathcal{L}_d^i(\theta_f, \theta_d) \right)$$

where \mathcal{L}_y and \mathcal{L}_d are the task and domain loss functions, respectively. Both n and n' are drawn from the dataset with the difference that n are training samples for which the task label is available, where for the n' samples only the domain is known. It is worth noting that n' are the samples with different domains in the training but can include samples from the test set, from which only the domain label is needed to adapt the shared feature representation. In Ganin et al. (2016) the authors find that such saddle points can be found using the following iterative stochastic gradient updates:

$$\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial \mathcal{L}_y^i}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_f} \right), \tag{4}$$

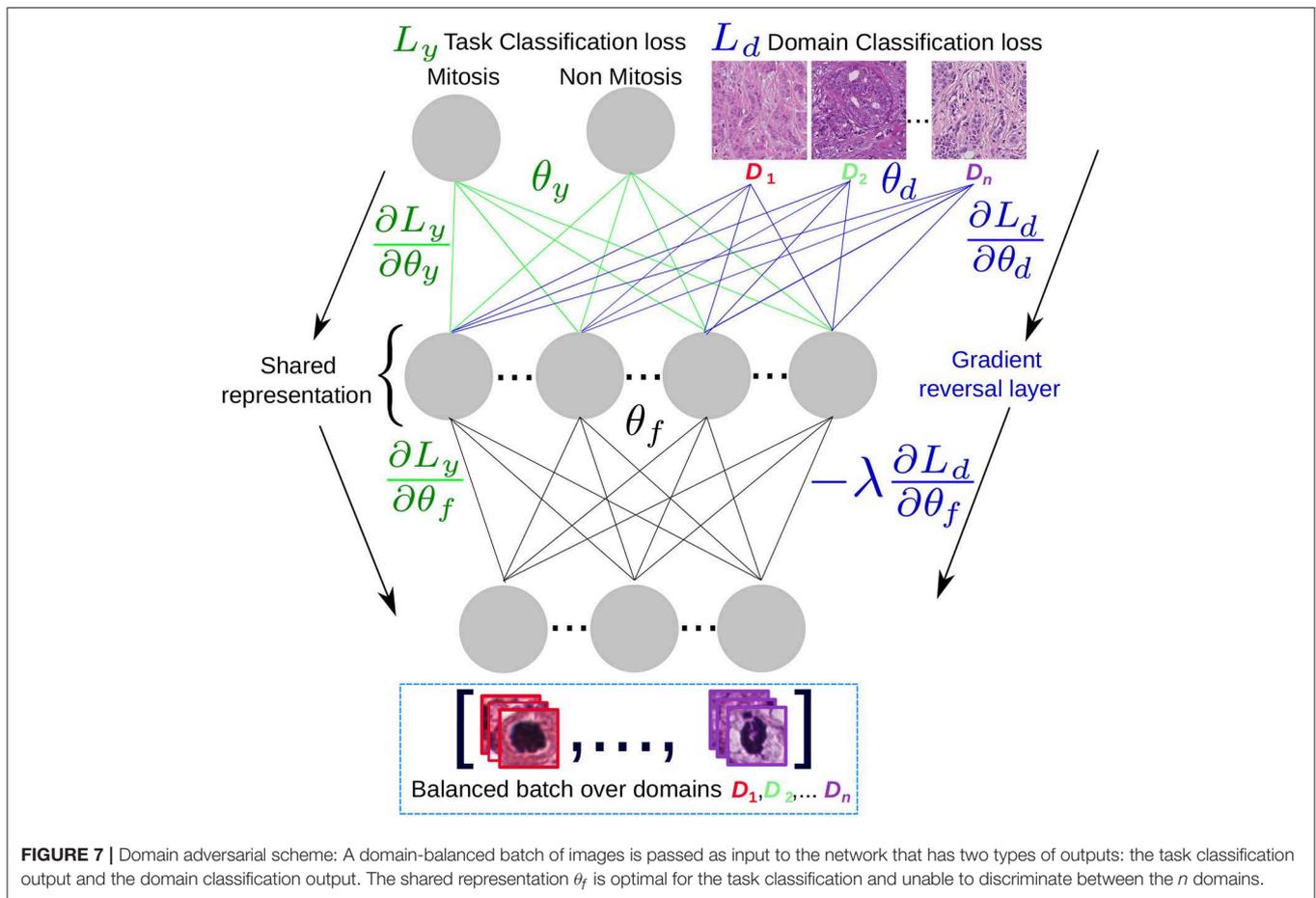
$$\theta_y \leftarrow \theta_y - \mu \frac{\partial \mathcal{L}_y^i}{\partial \theta_y}, \tag{5}$$

$$\theta_d \leftarrow \theta_d - \mu \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_d} \tag{6}$$

Where μ is the base learning rate for the task classifier and λ is the domain learning rate multiplier that allow to learn the domain-dependant features at a different pace from the task-related ones. Equations (5) and (6) optimize the parameters for the task and domain loss functions, respectively, as in a classical multi-output DCNN scenario. The loss for the task classifier is minimized while the loss for the domain classifier is maximized. This is done in Equation (4) that is optimized in an adversarial manner by going in the opposite gradient direction that minimizes task loss and in the positive direction of the gradient for the domain features, maximizing the domain classifier loss.

A simplified architecture scheme is shown in **Figure 7**, where the task and the domain classifiers are at the top of a shared representation that maximizes the classification performance to separate between mitotic and non-mitotic images and at the same time minimizes the probability to recover in which center (domain) the image was generated. If the model converges, then for a test image the inferred representation avoids to include any center-specific information. The original center for the test image could be unknown, since this label is not needed and we are only interested in obtaining the mitosis/non-mitosis probability.

The main characteristic of DANN models is that they do not require external models to perform staining normalization or color augmentation but are the task labels and the indicators from the scanners or domains that are used to drive the training process and solve the task without depending explicitly on such models. This feature is also a potential drawback if one is interested in a staining quantification scenario, because DANN models do not output an explicit normalized image but the normalization is done in the feature space instead. This group



samples from the same class regardless of the domain from where they were generated. Our experiments are inspired by the work of Lafarge et al. (2017) where the authors build a DCNN with a reverse gradient layer that aims to learn in an adversarial manner the mitotic probability of patches and their domain.

It was noted by Engilberge et al. (2017) that in the context of DCNN, the models are sensitive to color changes and also that in the more common architectures, the color-sensitive units are located in the first layers of the network, which suggests that the stain-dependent information of the network should be removed from the first layers of the architectures, this lead us to place the reversal gradient information before the fully connected layers of the network as shown in the experimental setup section 3.

3. EXPERIMENTAL SETUP

3.1. CNN Architectures

We designed a baseline DCNN model for both datasets. For the TUPAC dataset, we designed a relatively compact DCNN architecture due to the small input size of $96 \times 96 \times 3$ patches. The architecture is composed of two convolutional layers followed by two blocks of convolution, batch normalization, and max-pooling. Then, a dropout layer is included to regularize the model followed by another block of convolution, batch

normalization, and max-pooling. At the end of the network, two branches of dense layers are connected to the mitosis and domain class predictions. For the domain branch, the gradient is reversed in order to optimize (Equation 4).

For the TCGA-TMAZ dataset, we used as base model the MobilNet architecture used in the experiments by Arvaniti et al. (2018). The last layer was removed in order to add the two branches for the domain and Gleason pattern classification. The gradient reversal layer was added in the same way as in the TUPAC architecture, i.e., after the common convolutional filters and before the domain-specific dense features.

3.2. Implementation Details

We implemented the DCNN architectures and performed model training using the Keras deep learning framework with the Tensorflow backend. For the staining normalization method, we used the public implementation of the Macenko method from the StainTools library³. A crucial implementation detail is the need to generate domain-balanced batches for the adversarial update of DANN models; this is a dataset-specific part of the pipeline where the number of domains should be provided, since they are used as a one-hot encoding domain-label vector for each of the

³<https://github.com/Peter554/StainTools>

TABLE 3 | Results on the TUPAC dataset.

CNN model combinations	Baseline DCNN							
Color augmentation		✓				✓	✓	✓
Staining normalization			✓			✓		✓
Domain adversarial				✓			✓	✓
Internal test set (F1-score)	0.8088 (±0.02)	0.8117 (±0.001)	0.7630 (±0.04)	0.6950 (±0.379)	0.7787 (±0.03)	0.6985 (±0.01)	0.6945 (±0.02)	
External test set (F1-score)	0.71173 (±0.02)	0.7306 (±0.07)	0.5424 (±0.01)	0.8236 (±0.071)	0.5963 (±0.1)	0.6740 (±0.01)	0.5742 (±0.009)	
Internal test set (AUC)	0.9596 (±0.006)	0.9631 (±0.005)	0.9351 (±0.001)	0.8972 (±0.011)	0.9503 (±0.01)	0.9030 (±0.002)	0.8871 (±0.02)	
External test set (AUC)	0.8014 (±0.01)	0.8270 (±0.06)	0.848 (±0.075)	0.9146 (±0.003)	0.7925 (±0.06)	0.8446 (±0.004)	0.8255 (±0.06)	

Performance measures for the possible combinations of color augmentation, staining normalization, and DANN. The first column corresponds to the baseline DCNN without any staining normalization nor color augmentation. Numbers in bold indicate the best result for that row (performance measure).

TABLE 4 | Results for the TCGA-TMAZ dataset.

CNN model combinations	Baseline DCNN							
Color augmentation		✓				✓	✓	✓
Staining normalization			✓			✓		✓
Domain adversarial				✓			✓	✓
Internal test set (F1-score)	0.5614 (±0.01)	0.5386 (±0.03)	0.5928 (±0.05)	0.6232 (±0.03)	0.6761 (±0.01)	0.6493 (±0.01)	0.6317 (±0.01)	
External test set (F1-score)	0.4837 (±0.02)	0.5732 (±0.04)	0.5863 (±0.04)	0.5908 (±0.03)	0.6222 (±0.05)	0.5821 (±0.01)	0.5625 (±0.06)	
Internal test set (AUC)	0.8155 (±0.01)	0.7429 (±0.01)	0.7391 (±0.01)	0.8409 (±0.05)	0.7544 (±0.01)	0.7755 (±0.03)	0.7049 (±0.02)	
External test set (AUC)	0.6368 (±0.01)	0.6735 (±0.06)	0.6633 (±0.01)	0.6838 (±0.01)	0.6798 (±0.02)	0.6913 (±0.01)	0.6712 (±0.01)	

F1-scores for the possible combinations of color augmentation, staining normalization and domain adversarial. The first column corresponds to the baseline model. Numbers in bold indicate the best result for that row (performance measure).

samples. The learning rates were explored in the base models and fixed to $\mu = 0.01$ for the mitosis model and $\mu = 0.001$ for the Gleason pattern classification model, according to the best results on the test set.

A warmup of 100 batch iterations for the task branch ($\lambda = 0$) in the DANN model was observed to lead to a more stable training as compared to starting with random weights for both branches. We provide the code of our experiments for further implementation details⁴.

Four DCNN models were trained for each experiment combination, to minimize performance variations in the test sets due to the random-weight initializations of the DCNN models. Average and standard deviation in performance is the final result reported. In total 56 DCNN models were trained.

4. RESULTS

The classification results are shown in **Tables 3, 4**. Each of the cells shows the average performance of the four DCNN initializations and standard deviation in parenthesis. Each column represents a combination of the strategies, and the first column is the performance of the base DCNN model without any staining normalization or augmentation technique. We trained the DCNN models without dropout but observed overfitting and degraded performance on the test sets. Therefore, all of our baseline DCNN include dropout with a probability of 0.25 in the

layers described in section 3. The F1 score was selected as one of the performance measures to account for the class imbalance of our datasets. The other measure selected was the area under the receiver operating characteristics curve (AUC) since the binary decision threshold is not always close to 0.5 for all the models. Qualitative UMAP visualizations of feature embeddings for the external TUPAC dataset are displayed in **Figure 8**.

4.1. TUPAC

In the TUPAC dataset experiments, color augmentation shows a good performance on the internal test set, while the performance on the external test set is slightly above the baseline (**Table 3**). This result might be due to variance in the external test set staining that can not be captured by color augmentation alone. When both color augmentation and DANN are combined, the performance is not outstanding neither for internal nor external datasets. However, the results represent a good tradeoff, close to the best results for the external test dataset. The best results for the external dataset are obtained with the DANN approach. An interesting result, yet not fully understood, is that DANN show an essential decrease in performance on the internal test set. Staining normalization decreased the performance in both the internal and external test set images, likely due to overfitting in the training domains.

4.2. TCGA-TMAZ

For this dataset the best results involve color augmentation and DANN. DANN alone obtained the best AUC performance

⁴https://github.com/sebastianffx/stain_adversarial_learning

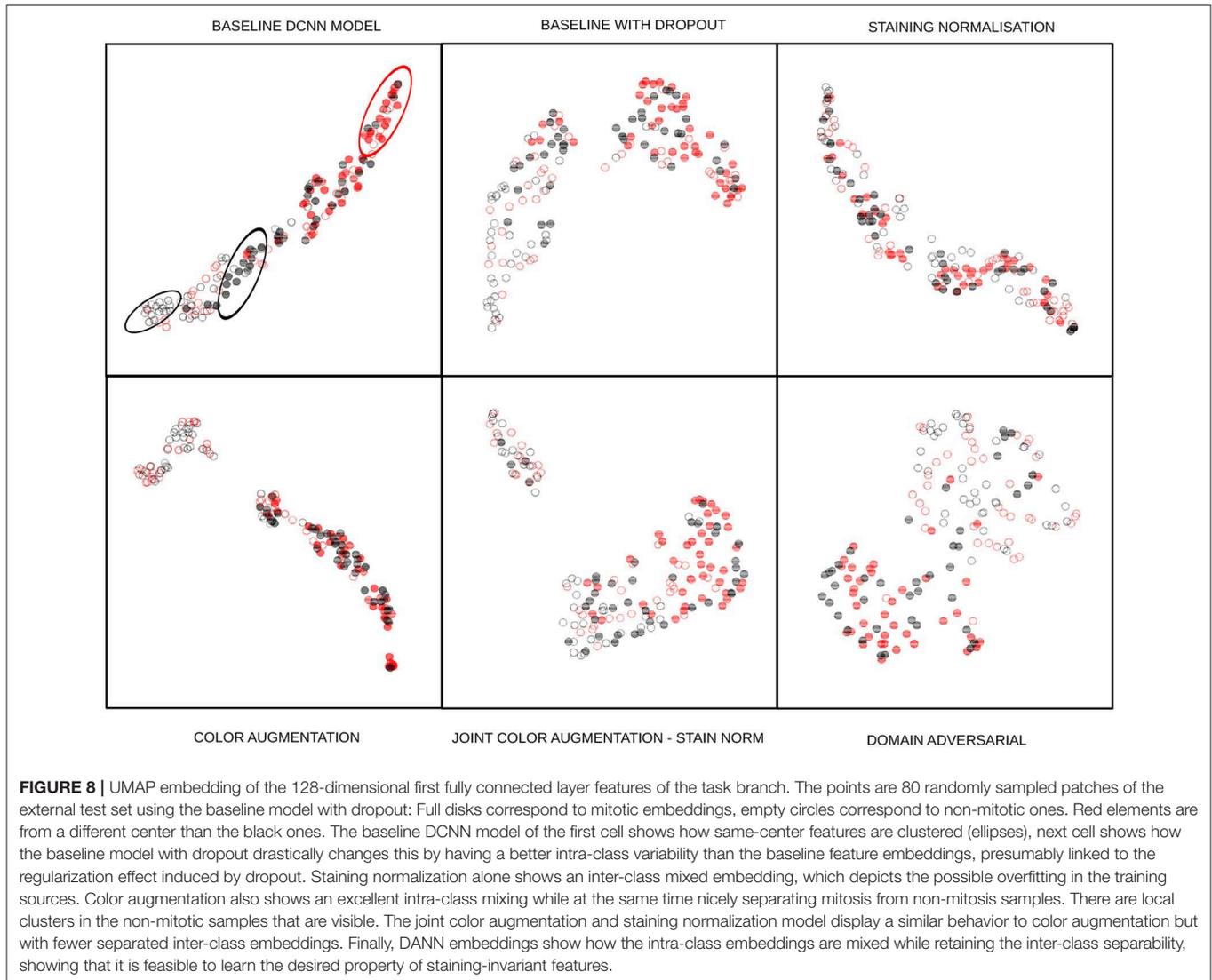


FIGURE 8 | UMAP embedding of the 128-dimensional first fully connected layer features of the task branch. The points are 80 randomly sampled patches of the external test set using the baseline model with dropout: Full disks correspond to mitotic embeddings, empty circles correspond to non-mitotic ones. Red elements are from a different center than the black ones. The baseline DCNN model of the first cell shows how same-center features are clustered (ellipses), next cell shows how the baseline model with dropout drastically changes this by having a better intra-class variability than the baseline feature embeddings, presumably linked to the regularization effect induced by dropout. Staining normalization alone shows an inter-class mixed embedding, which depicts the possible overfitting in the training sources. Color augmentation also shows an excellent intra-class mixing while at the same time nicely separating mitosis from non-mitosis samples. There are local clusters in the non-mitotic samples that are visible. The joint color augmentation and staining normalization model display a similar behavior to color augmentation but with fewer separated inter-class embeddings. Finally, DANN embeddings show how the intra-class embeddings are mixed while retaining the inter-class separability, showing that it is feasible to learn the desired property of staining-invariant features.

on the internal test set and DANN combined with color augmentation obtained the best AUC performance for the external test. Staining normalization and color augmentation together obtained the best F1 scores, both for the internal and external datasets.

4.3. Statistical Significance of the Results

Comparing the decisions of the best vs. second-best performing methods using the Wilcoxon signed-rank test yield the following results: For the TUPAC dataset, the difference between the color augmentation and DANN, were compared only for the external dataset, since the number of patches in the internal test set always led to a p -value of 0. In the external test set, the difference was significant ($p < 0.05$) in two out of four runs: $p < 1.7232 \times 10^{-7}$, 0.0723, 0.1808, and 1.8721×10^{-9} , respectively.

For the TCGA-TMA dataset, the difference between the combination of color augmentation and staining normalization vs. the combination of DANN and color augmentation was

significant ($p < 0.05$) in three out of four runs for the internal test dataset ($p < 1.0922 \times 10^{-5}$, 0.1919, 0.0008, 1.5824×10^{-6} , respectively), and significant in the four runs for the external test dataset ($p < 1.9029 \times 10^{-11}$, 6.0386×10^{-20} , 1.0652×10^{-32} , 3.1229×10^{-23}).

5. DISCUSSION AND CONCLUSIONS

The results show the importance of having separate evaluation datasets, as the results in only one can be misleading. Statistical significance of the results suggests that DANN and color augmentation (also in combination with staining normalization) can deal with the problem of having a limited set of centers for training deep learning models.

In the TCGA-TMAZ dataset, the performance gain of DANN+CA with respect to the baseline is considerable and shows how naively training DCNN models without any strategy to overcome staining variety is suboptimal.

The combination of color augmentation, staining normalization, and DANN training of DCNN models did not improve the results and, in some cases, obtained results below the baseline. Such behavior might be due to overfitting to the train centers because color augmentation over the normalized images accounts only for a limited range of variation (those in the training centers) and leading to domain adversarial training to not learn enough staining invariances. Similar behavior occurs when training DANN with stain-normalized images, for which the results are not reported here in the table, but were close and in some cases worse than the baseline.

To conclude, the experimental results show that staining normalization, color augmentation, and DANN methods improve DCNN generalization for classification tasks using digital pathology images. Results did not show a clear *winner* or combination strategy. Statistical significance tests of the results suggest that the use of color augmentation can alleviate color heterogeneity problems up to some extent and that DANN training of DCNN models alone or in combination with color augmentation can lead to even better results.

Designing deep learning experiments for computational pathology with images from different centers can provide meaningful insights about the performance of the classification algorithms in realistic scenarios (for instance by predicting class labels for data with the same pathology but scanned under different staining conditions). In future work, we devise

a training of DANN with staining augmentation done in an end-to-end architecture, also evaluating DANN performance thoroughly, including the external set images in the fine-tuning of the model.

DATA AVAILABILITY

The datasets analyzed for this study can be found in the TCGA-PRAD repository (<https://portal.gdc.cancer.gov/projects/TCGA-PRAD>) and the Replication Data for Automated Gleason grading of prostate cancer tissue microarrays via deep learning (doi: 10.7910/DVN/OCYCMP).

AUTHOR CONTRIBUTIONS

SO, MA, VA, AK, and HM conceived the study and wrote the manuscript. SO carried out the experiments.

FUNDING

This work described in this paper has received partial funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825292 (ExaMode). SO thanks Colciencias through the call 756 for PhD studies.

REFERENCES

- Al-Janabi, S., Huisman, A., and Van Diest, P. J. (2012). Digital pathology: current status and future perspectives. *Histopathology* 61, 1–9. doi: 10.1111/j.1365-2559.2011.03814.x
- Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., et al. (2018). Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Sci. Rep.* 8:12054. doi: 10.1038/s41598-018-30535-1
- Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermesen, M., et al. (2019). From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Trans. Med. Imaging* 38, 550–560. doi: 10.1109/TMI.2018.2867350
- Bejnordi, B. E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., et al. (2016). Stain specific standardization of whole-slide histopathological images. *IEEE Trans. Med. Imaging* 35, 404–415. doi: 10.1109/TMI.2015.2476509
- Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318, 2199–2210. doi: 10.1001/jama.2017.14585
- Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). “Analysis of representations for domain adaptation,” in *Advances in Neural Information Processing Systems*, eds B. Schölkopf, J. C. Platt, and T. Hoffman (Vancouver, BC: MIT Press), 137–144.
- Bentaieb, A. and Hamarneh, G. (2018). Adversarial stain transfer for histopathology image analysis. *IEEE Trans. Med. Imaging* 37, 792–802. doi: 10.1109/TMI.2017.2781228
- Bug, D., Schneider, S., Grote, A., Oswald, E., Feuerhake, F., Schüller, J., et al. (2017). “Context-based normalization of histological stains using deep convolutional features,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, eds J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. S. Tavares, M. Moradi, et al. (Québec City, QC: Springer), 135–142.
- Ciampi, F., Geessink, O., Bejnordi, B. E., de Souza, G. S., Baidoshvili, A., Litjens, G., et al. (2017). “The importance of stain normalization in colorectal tissue classification with convolutional networks” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne: IEEE), 160–163.
- Crammer, K., Kearns, M., and Wortman, J. (2006). “Learning from data of variable quality,” in *Advances in Neural Information Processing Systems*, eds Y. Weiss, B. Schölkopf, and J. C. Platt (Vancouver, BC: MIT Press), 219–226.
- Cruz-Roa, A., Gilmore, H., Basavanahally, A., Feldman, M., Ganesan, S., Shih, N., et al. (2018). High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: application to invasive breast cancer detection. *PLoS ONE* 13:e0196828. doi: 10.1371/journal.pone.0196828
- Engilberge, M., Collins, E., and Süsstrunk, S. (2017). “Color representation in deep neural networks,” in *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing: IEEE), 2786–2790.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.1007/978-3-319-58347-1_10
- Janowczyk, A., Basavanahally, A., and Madabhushi, A. (2017). Stain normalization using sparse autoencoders (stanosa): application to digital pathology. *Comput. Med. Imaging Graph.* 57, 50–61. doi: 10.1016/j.compmedimag.2016.05.003
- Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* 7:29. doi: 10.4103/2153-3539.186902
- Kothari, S., Phan, J. H., Stokes, T. H., Osunkoya, A. O., Young, A. N., and Wang, M. D. (2014). Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J. Biomed. Health Inform.* 18, 765–772. doi: 10.1109/JBHI.2013.2276766
- Kothari, S., Phan, J. H., Stokes, T. H., and Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *J. Am. Med. Inform. Assoc.* 20, 1099–1108. doi: 10.1136/amiainjnl-2012-001540
- Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., Moeskops, P., and Veta, M. (2017). “Domain-adversarial neural networks to address the appearance variability of histopathology images,” *Lecture Notes in Computer Science (Including Subseries*

- Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (Cham), 10553. doi: 10.1007/978-3-319-67558-9-10
- Leo, P., Lee, G., Shih, N. N., Elliott, R., Feldman, M. D., and Madabhushi, A. (2016). Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J. Med. Imaging* 3:047502. doi: 10.1117/1.JMI.3.4.047502
- Li, X., Member, S., and Plataniotis, K. N. (2015). A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. *IEEE Trans. Biomed. Eng.* 62, 1862–1873. doi: 10.1109/TBME.2015.2405791
- Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., et al. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* 6:26286. doi: 10.1038/srep26286
- Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., et al. (2009). “A method for normalizing histology slides for quantitative analysis,” in *2009. ISBI'09. IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (Boston, MA: IEEE), 1107–1110.
- Madabhushi, A. (2009). Digital pathology image analysis: opportunities and challenges. *Imaging Med.* 1, 7–10. doi: 10.2217/IIM.09.9
- Madabhushi, A., Agner, S., Basavanahally, A., Doyle, S., and Lee, G. (2011). Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Comput. Med. Imaging Graph.* 35, 506–514. doi: 10.1016/j.compmedimag.2011.01.008
- McCann, M. T., Majumdar, J., Peng, C., Castro, C. A., and Kovačević, J. (2014). “Algorithm and benchmark dataset for stain separation in histology images,” in *2014 IEEE International Conference on Image Processing (ICIP)* (Paris: IEEE), 3953–3957.
- Nagpal, K., Foote, D., Liu, Y., Wulczyn, E., Tan, F., Olson, N., et al. (2018). Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *arXiv arXiv:1811.06497*.
- Ren, J., Hacihaliloglu, I., Singer, E. A., Foran, D. J., and Qi, X. (2018). “Adversarial domain adaptation for classification of prostate histopathology whole-slide images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention Medical Image Computing and Computer Assisted Intervention - MICCAI 2018* (Granada), 201–209.
- Rousson, M., Hedlund, M., Andersson, M., Jacobsson, L., Låthén, G., Norell, B., et al. (2018). “Tumor proliferation assessment of whole slide images,” in *Medical Imaging 2018: Digital Pathology*, Vol. 10581, eds J. E. Tomaszewski and M. N. Gurcan (Houston, TX: SPIE - International Society For Optics and Photonics), 105810Y.
- Roy, S., kumar Jain, A., Lal, S., and Kini, J. (2018). A study about color normalization methods for histopathology images. *Micron* 114, 42–61. doi: 10.1016/j.micron.2018.07.005
- Shaban, M. T., Baur, C., Navab, N., and Albarqouni, S. (2018). “StainGAN: stain style transfer for digital histological images,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)* (Venice), 1–8.
- Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J., and Ciompi, F. (2018a). “H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection,” in *Medical Imaging 2018: Digital Pathology*, Vol. 10581, eds J. E. Tomaszewski and M. N. Gurcan (Houston, TX: International Society for Optics and Photonics), 105810Z.
- Tellez, D., Balkenhol, M., Otte-Holler, I., van de Loo, R., Vogels, R., Bult, P., et al. (2018b). Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging* 37, 2126–2136. doi: 10.1109/TMI.2018.2820199
- Tellez, D., Litjens, G., Bandi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *arXiv arXiv:1902.06543*.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ed E. Mortensen (Honolulu), 7167–7176.
- Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., et al. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* 35, 1962–1971. doi: 10.1109/TMI.2016.2529665
- Van Eycke, Y.-R., Allard, J., Salmon, I., Debeir, O., and Decaestecker, C. (2017). Image processing in digital pathology: an opportunity to solve inter-batch variability of immunohistochemical staining. *Sci. Rep.* 7:42964. doi: 10.1038/srep42964
- Van Eycke, Y.-R., Balsat, C., Verset, L., Debeir, O., Salmon, I., and Decaestecker, C. (2018). Segmentation of glandular epithelium in colorectal tumours to automatically compartmentalise ihc biomarker quantification: a deep learning approach. *Med. Image Anal.* 49, 35–45. doi: 10.1016/j.media.2018.07.004
- Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., et al. (2019). Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Med. Image Anal.* 54, 111–121. doi: 10.1016/j.media.2019.02.012
- Veta, M., Van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., et al. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.* 20, 237–248. doi: 10.1016/j.media.2014.11.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Otálora, Atzori, Andrearczyk, Khan and Müller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.