Medical Image Analysis 84 (2023) 102702

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-site cross-organ calibrated deep learning (MuSClD): Automated diagnosis of non-melanoma skin cancer☆

Yufei Zhou [a], Can Koyuncu [b,c], Cheng Lu [b], Rainer Grobholz [d,e], Ian Katz [f,g,2], Anant Madabhushi [b,h,4,1,*], Andrew Janowczyk [b,i,j,3]

[a] Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, USA
[b] Department of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA, USA
[c] Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, USA
[d] Institute of Pathology, Cantonal Hospital Aarau, Aarau, Switzerland
[e] Medical Faculty University of Zurich, Zurich, Switzerland
[f] Southern Sun Pathology, Sydney, NSW, Australia
[g] University of Queensland, Brisbane, Qld, Australia
[h] Atlanta VA Medical Center, Atlanta, USA
[i] Department of Oncology, Lausanne University Hospital
[j] Department of Diagnostics, Division of Clinical Pathology, Geneva University Hospitals

ARTICLE INFO

Keywords:
Non-melanoma skin cancer
Subtyping
Deep learning
Digital pathology
Domain adaptation
Domain shift

ABSTRACT

Although deep learning (DL) has demonstrated impressive diagnostic performance for a variety of computational pathology tasks, this performance often markedly deteriorates on whole slide images (WSI) generated at external test sites. This phenomenon is due in part to domain shift, wherein differences in test-site pre-analytical variables (e.g., slide scanner, staining procedure) result in WSI with notably different visual presentations compared to training data. To ameliorate pre-analytic variances, approaches such as CycleGAN can be used to calibrate visual properties of images between sites, with the intent of improving DL classifier generalizability. In this work, we present a new approach termed Multi-Site Cross-Organ Calibration based Deep Learning (MuSClD) that employs WSIs of an *off-target* organ for calibration created at the same site as the *on-target* organ, based off the assumption that cross-organ slides are subjected to a common set of pre-analytical sources of variance. We demonstrate that by using an off-target organ from the test site to calibrate training data, the domain shift between training and testing data can be mitigated. Importantly, this strategy uniquely guards against potential data leakage introduced during calibration, wherein information only available in the testing data is imparted on the training data. We evaluate MuSClD in the context of the automated diagnosis of non-melanoma skin cancer (NMSC). Specifically, we evaluated MuSClD for identifying and distinguishing (a) basal cell carcinoma (BCC), (b) in-situ squamous cell carcinomas (SCC-In Situ), and (c) invasive squamous cell carcinomas (SCC-Invasive), using an Australian (training, $n = 85$) and a Swiss (held-out testing, $n = 352$) cohort. Our experiments reveal that MuSCID reduces the Wasserstein distances between sites in terms of color, contrast, and brightness metrics, without imparting noticeable artifacts to training data. The NMSC-subtyping performance is statistically improved as a result of MuSCID in terms of one-vs. rest AUC: BCC (0.92 vs 0.87, $p = 0.01$), SCC-In Situ (0.87 vs 0.73, $p = 0.15$) and SCC-Invasive (0.92 vs 0.82, p = 1e-5). Compared to baseline NMSC-subtyping with no calibration, the

☆ This article is part of a collaborative joint Special Issue between Medical Image Analysis and IEEE Transactions on Medical Imaging to publish high quality research on COVID-19.

* Corresponding author.
E-mail address: anantm@emory.edu (A. Madabhushi).

[1] Dr. Madabhushi is an equity holder in Picture Health, Elucid Bioimaging, and Inspirata Inc. Currently, he serves on the advisory board of Picture Health, Aiforia Inc, and SimBioSys. He also currently consults for Biohme, SimBioSys, and Castle Biosciences. He also has sponsored research agreements with AstraZeneca, Boehringer-Ingelheim, Eli-Lilly and Bristol Myers-Squibb. His-technology has been licensed to Picture Health and Elucid Bioimaging. He is also involved in 3 different R01 grants with Inspirata Inc.

[2] Dr. Katz has an options equity stake in PathologyWatch as a Medical advisor.

[3] Dr. Janowczyk is involved with an InnoSuisse grant with Lunaphore, and provides consulting for Merck, Roche, and Lunaphore.

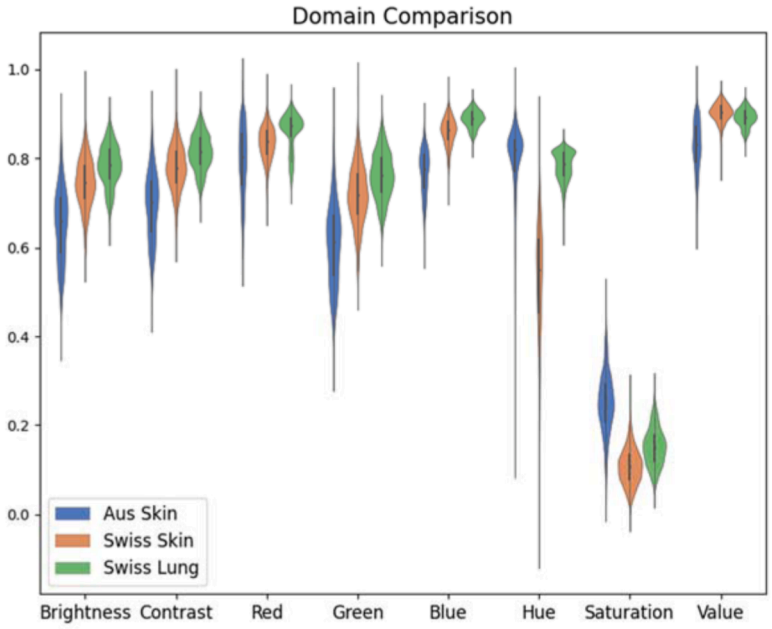
[4] Denotes co-Senior Authorship.

https://doi.org/10.1016/j.media.2022.102702
Received 4 January 2022; Received in revised form 9 November 2022; Accepted 21 November 2022
Available online 24 November 2022

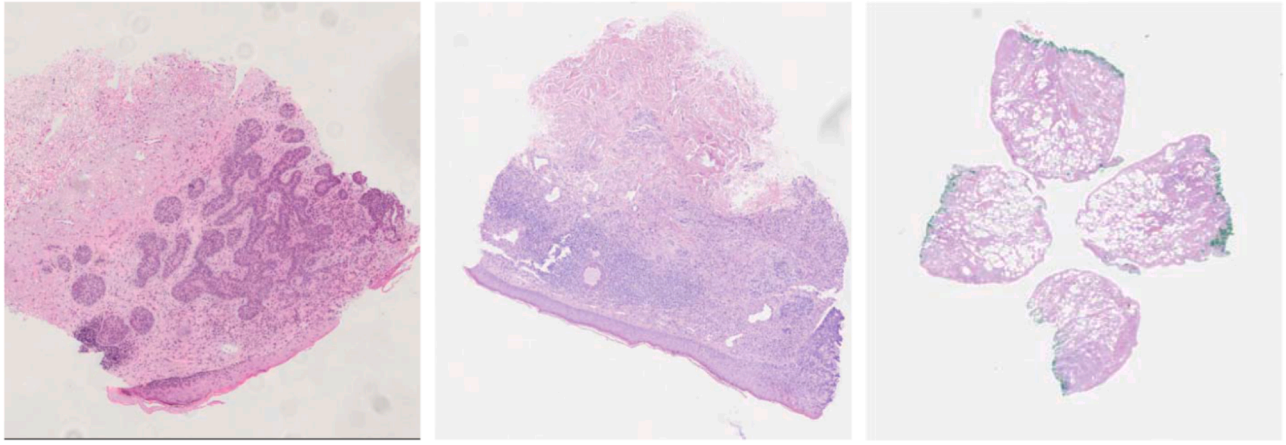
internal validation results of MuSClD (BCC (0.98), SCC-In Situ (0.92), and SCC-Invasive (0.97)) suggest that while domain shift indeed degrades classification performance, our on-target calibration using off-target tissue can safely compensate for pre-analytical variabilities, while improving the robustness of the model.

1. Introduction

Globally, there has been a significant increase in non-melanoma skin cancer (NMSC) incidence, especially in Caucasians (Samarasinghe and Madan, 2012). Seventy-five percent of NMSC cases correspond to basal cell carcinoma (BCC) which has a low risk of mortality (<0.1%) (Samarasinghe and Madan, 2012). The majority of the remaining NMSC cases are squamous cell carcinomas (SCC), which when left untreated are far more likely to metastasize to other organs (0.3–3.7%), leading to an increased risk of mortality (Samarasinghe and Madan, 2012). As such, an accurate differential diagnosis between BCC and SCC remains critical. SCC can be either SCC-In Situ (Bowen disease) or SCC-Invasive (Yanofsky et al., 2011). SCC-In Situ is a superficial form of SCC; however, it has relatively high risk (3%−5%) of progression to SCC-Invasive



(a)



(b) (c) (d)

Fig. 1. The violin plot (a) illustrates the inter-site domain shift between Australian/Swiss skin cohorts in terms of brightness (mean intensity of each color channel), contrast, hue/saturation/value (HSV), and red/green/blue (RGB) values. Interestingly, the shift between the skin and lung cohorts from Switzerland (Swiss) is less pronounced, likely as a result of similar pre-analytic variables (e.g., staining protocol, scanner, tissue sectioning, and slide preparation factors). This similarity suggests the feasibility of using Swiss lung data as calibration templates for Australian skin slides. BCC cases from the (b) Australian and (c) Swiss skin cohorts show that Swiss skin slides are generally bluer, a concept quantitatively reflected in (a). Interestingly, (d) a lung sample from the Swiss site has a similar visual appearance to (c) the Swiss skin image, a notion again supported by the image metrics in (a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

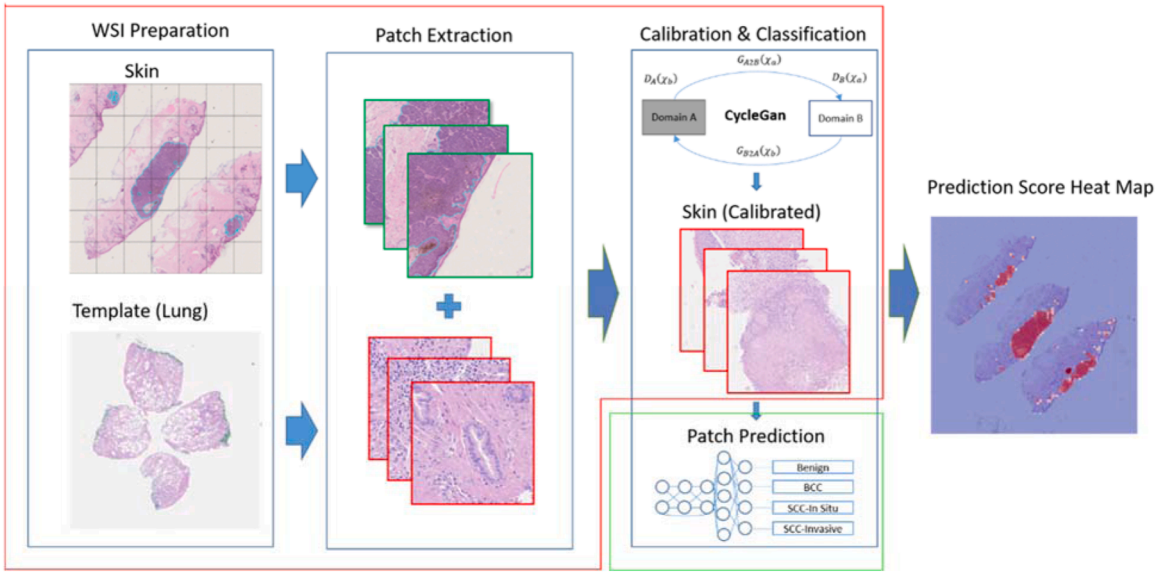
(Yanofsky et al., 2011). Since approximately 1.1% of women and 2.4% of men with SCC-Invasive eventually develop tumor metastases, accurate diagnosis and close monitoring of SCC subtypes is warranted (Venables et al., 2019).

The diagnosis of NMSC is performed by pathologists identifying characteristic histological features of each NMSC subtype from hematoxylin and eosin (H&E) stained tissue specimens (Brown et al., 1979). BCC is a type of basaloid epithelial tumor arising from the basal layer of the epidermis (Elder et al., 2018; Mackiewicz-Wysocka et al., 2013). The basaloid cells form a regular palisade at the periphery of the tumor nest while their distribution in the middle of the nest is chaotic (Elder et al., 2018; Mackiewicz-Wysocka et al., 2013). Necrotic bodies may develop in the center of the tumor nest leading to formation of cystic spaces containing mucinous material (Elder et al., 2018; Mackiewicz-Wysocka et al., 2013). One key feature distinguishing BCC from other basaloid tumors is the surrounding dense stroma with mucinous material (Elder et al., 2018; Mackiewicz-Wysocka et al., 2013). Another diagnostic feature of BCC is its empty peritumoral cleft due to the retraction of stroma (Elder et al., 2018; Mackiewicz-Wysocka et al., 2013). On the other hand, SCC is often characterized by its relatively large polyhedral cells with abundant, glassy eosinophilic cytoplasm with copious keratin formation (Elder et al., 2018). While the consensus among dermatopathologists tends to be high in identifying BCC (>95%), concordance is relatively lower for SCC (77%) (Onajin et al., 2015). In addition, it may be non-trivial to distinguish the subtypes of SCC in some cases, specifically SCC-In Situ versus SCC-Invasive. The development and application of computerized digital pathology solutions to distinguish between BCC vs SCC and also SCC-in Situ vs SCC-Invasive would meaningfully aid in providing improved diagnosis, prognosis, and treatment management of patients with NMSC (Jiang et al., 2019; Marka et al., 2019).

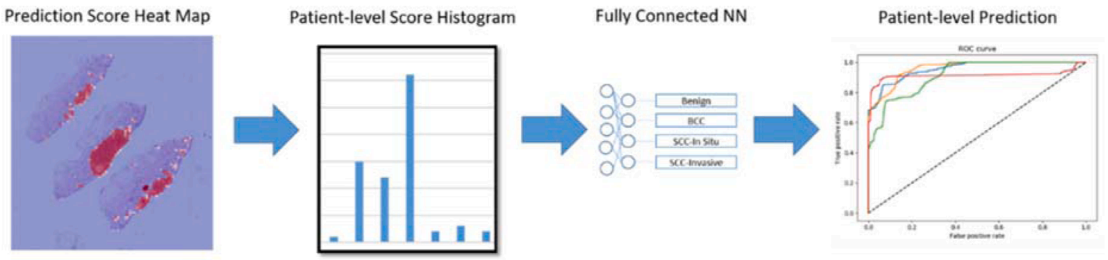
Deep learning (DL) strategies are well suited to this multiclass classification subtyping task and can enable the improved identification of benign tissue, BCC, and SCC (Jiang et al., 2019; Kimeswenger et al., 2020; Marka et al., 2019; Wang et al., 2019). Unfortunately, there is mounting evidence that DL performance is brittle in the context of pre-analytical variabilities in slide preparation (e.g., staining protocol, slide scanner, tissue thickness), and we show that NMSC is not exempt from these issues (Shaban et al., 2019; Tellez et al., 2019). Each source of pre-analytic variance has a unique and additive impact both on the presentation of the tissue slide and its associated whole slide image (WSI) (Shaban et al., 2019). For example, there are visually perceivable differences in terms of image hue/saturation/value (HSV), red/blue/-green (RGB), contrast, and brightness values between similar BCC cases from two different sites (see Fig. 1(b) and (c)). These variabilities contribute to a phenomenon known as "***domain shift***" (Stacke et al., 2019), wherein testing data ($\boldsymbol{D_V}$) lies on a different underlying distribution as compared to training data ($\boldsymbol{D_T}$). These differences in the underlying image distribution between $\boldsymbol{D_T}$ and $\boldsymbol{D_V}$ have been shown to heavily impact DL performance (Jiang et al., 2019; Tellez et al., 2019). This performance drop is exacerbated when deploying a trained DL model to a new site, where pre-analytical differences, and thus domain shift, are often most substantial.

To compensate for domain shift, several domain adaptation techniques have been proposed that aim to "***calibrate***" the data from a source domain to a target domain, helping to ameliorate pre-analytic differences using post-image acquisition processing steps (e.g., stain normalization) (de Bel et al., 2019; Ganin and Lempitsky, 2014; Shaban et al., 2019; Tellez et al., 2019). More recent approaches have employed CycleGAN, a type of generative adversarial network (GANs) (de Bel et al., 2019; Guha et al., 2020; Shaban et al., 2019; Tellez et al., 2019; You et al., 2020a). In these studies, "***template***" images are often sampled from the ***target*** domain (e.g., testing site) as the reference to calibrate the ***source*** domain data (e.g., training site). Growing evidence (Bel et al., 2021; Shaban et al., 2019) suggests that CycleGAN achieves superior performance in mitigating the impact of domain shift on downstream image analysis tasks as compared to handcrafted stain normalization approaches (Macenko et al., 2009; Reinhard et al., 2001). However, even current approaches for addressing domain shift suffer from two potential issues in terms of introducing artifacts and experimental design concerns. In the first case, aligning $\boldsymbol{D_V}$ to the $\boldsymbol{D_T}$ provides an opportunity for the introduction of artifacts including blur, checkerboard artifacts (Aitken et al., 2017; Zhang et al., 2019), or texture distortions. These artifacts, as well as compression artifacts, have generally been shown to have a negative impact on the performance of DL image analysis pipelines (Chen et al., 2020; Dodge and Karam, 2016), for instance in a recent DL-based tumor-stroma ratio algorithm (Foucart et al., 2018; Wright et al., 2021). To avoid this performance degradation on the $\boldsymbol{D_V}$, other approaches instead involve calibrating *training* images to more closely resemble the anticipated $\boldsymbol{D_V}$ properties. This process yields a site-specific DL model, which can be directly applied to new test images, mitigating the possibility of artifact introduction. However, these approaches typically employ *testing* images *from the same organ and task* (i.e., on-target organ) as templates for $\boldsymbol{D_T}$ calibration (de Bel et al., 2019; Tellez et al., 2019). In such *same-organ calibration (SOC)* setups, *data leakage* may take place wherein the calibration model learns and then imparts information from the held-out $\boldsymbol{D_V}$ into the $\boldsymbol{D_T}$, in other words violating the strict separation between $\boldsymbol{D_T}$ and $\boldsymbol{D_V}$. Such leakage may subsequently inflate the model testing accuracy, degrading the generalizability of the model (Chiavegatto Filho et al., 2021; Dong, 2022; Kaufman et al., 2012; Tampu et al., 2022). For example, a CycleGAN could impart the *task-specific knowledge* that BCC cells from an external test site are slightly larger, due to microns per pixel differences in the scanner, into training images by modifying their size. This then begs the question of whether the external site data can still be considered an independent $\boldsymbol{D_V}$ for assessing the robustness of the machine classifier, now that the classifier has been inadvertently exposed to key attributes of the $\boldsymbol{D_V}$. It was also reported (Wei et al., 2019) that a CycleGAN model can easily render visual attributes of precancerous tissue onto normal tissue inputs, wherein the CycleGAN learned and transferred task-specific features from precancerous tissue templates to the training images. Moreover, Dong et al. suggest only preprocessing training data to prevent data leakage, therefore calibration of $\boldsymbol{D_T}$ rather than $\boldsymbol{D_V}$ is also in favor of reducing data leakage risk (Dong, 2022). Taken together, it stands to reason that a superior calibration approach could help disentangle and thereby learn site-specific pre-analytic variables, while *being blinded from task-specific information*, potentially contaminating classifier construction.

We hypothesize that site-specific pre-analytic variables imparted into WSI are sufficiently similar between organs, such that images from a second "*off-target*" organ (i.e., an organ not employed in the training and testing of a corresponding diagnosis task) can be used as a template for calibration of the primary on-target organ, thus yielding performance improvements for the target task. The usage of off-target organs for on-target calibration is thus termed "*cross-organ*" calibration. To evaluate this hypothesis, we measure improvement in the performance of a DL-based non-melanoma skin cancer (NMSC) subtyping classifier after calibrating the skin training images with lung template images, an approach we term Multi-site Cross-organ Calibration Deep Learning (MuSClD). Our subtyping classifier is trained using an Australian cohort ($n = 85$) to distinguish between different subtypes of non-melanoma skin cancer: benign, basal cell carcinoma (BCC), in-situ squamous cell carcinoma (SCC), and invasive SCC. This NMSC-subtyping model is subsequently evaluated on an independent Swiss cohort ($n = 352$). Lung tissue samples from the Swiss site were employed as templates for MuSCID of the Australian data, in order to help mitigate domain shift effects. This lung tissue (see Fig. 1(d)) naturally shares similar variables associated with laboratory equipment (e.g., microtome, scanner, stainer) and biochemical properties (temperature, humidity, stain), resulting in similar image characteristics to that of skin samples (see Fig. 1(a)). Additionally, since the lung and skin tissue are unrelated to each other, the use of cross-organ information helps mitigate the possibility of *data leakage*. We also demonstrate that despite the differences



(a)



(b)

Fig. 2. (a) Overview of an example MuSCID (red box) being applied to an NMSC subtyping pipeline. Patches with a size of 512 × 512 at 20x magnification are obtained in the calibration stage (using CycleGAN model in the red box) as the input to the NMSC subtype classification network (using ResNext50_32 × 4d model in the green box). Tumor localization is visualized with a heat-map of patch prediction scores. (b) To obtain the WSI-level prediction, we collect the output probability of the predicted subtype of all patches in a WSI to form a patient-level subtype histogram. A simple fully-connected neural network was trained with these subtype histograms to predict the patient-level diagnosis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Notation table for sites and organs.

Notation	Definition
A	Training Site
B	Testing Site
S	Skin Tissue
L	Lung Tissue
A_S	Skin Slides from site A
A_S^L	Skin Slides from site A calibrated with Lung Templates (MuSCID)
A_S^S	Skin slides from site A calibrated with Skin Templates (SOC)
B_S	Skin Slides from B
B_L	Lung Slides from B
G	Generator networks in CycleGAN
D	Discriminator networks in CycleGAN
M_C	Downstream classification model trained with calibrated data.
M_N	Downstream classification model trained with un-calibrated data.

in lung and skin tissue morphology, the fidelity of calibration outputs is retained.

2. Related works and novel contributions

Previous works in the NMSC space have mainly focused on the identification and classification of either BCC alone or BCC vs. SCC-Invasive. A recent review provides an excellent summary of 39 machine learning-based NMSC detection algorithms, spanning both DL and non-DL-based methods (Marka et al., 2019). Kimeswenger et al. reported a DL-based model that was able to identify BCC lesions from WSI-level inputs, while Jiang et al. reported a similar application of the DL-based model not only in WSI but also in microscope ocular images (MOI) captured by smartphones (Jiang et al., 2019; Kimeswenger et al., 2020). Although this collection of works mostly focused on the

Table 2
Composition of Cohorts. WSIs in A were scanned at a magnification of 20x. WSIs in B were scanned at a 40x magnification and down-sampled to 20x during pre-processing to approximate the resolution of slides in A.

	Training Cohort from site A (Australia)			Testing Cohort from site B (Switzerland)	
	# Patients	# WSI	# Patches	# Patients	# WSI
Benign	11	13	7905	9	9
BCC	54	70	2652	131	131
SCC-In Situ	10	12	3132	12	12
SCC-Invasive	10	11	2216	200	200
Total	85	106	15,905	352	352

Table 3
Description of comparison strategies. The letter E represents an experiment with the first subscript indicating the $\boldsymbol{D_T}$ and the second describing the corresponding $\boldsymbol{D_V}$. A_S, A_S^L, A_S^S and B_S denote the original Australian skin data, the calibrated Australian skin (by Swiss lung), calibrated Australian skin (by Swiss skin), and the held-out $\boldsymbol{D_V}$ of Swiss skin, respectively.

Comparison Strategy	D_T	D_V	Network Model Used
$\boldsymbol{E_{A_S,\ A_S}}$	A_S	A_S	Train from scratch.
$\boldsymbol{E_{A_S^L,\ A_S^L}}$	A_S^L	A_S^L	Train from scratch.
$\boldsymbol{E_{A_S,\ A_S^L}}$	A_S	A_S^L	Reuse $E_{A_S,\ A_S}$
$\boldsymbol{E_{A_S^L,\ A_S}}$	A^L	A	Reuse $E_{A_S^L,\ A_S^L}$
$\boldsymbol{E_{A_S^S,\ A_S^S}}$	A^S	A^S	Train from scratch.
$\boldsymbol{E_{A_S^L,\ B_S}}$	A_S^L	B_S	Reuse $E_{A_S^L,\ A_S^L}$
$\boldsymbol{E_{A_S^S,\ B_S}}$	A_S^S	B_S	Reuse $E_{A_S^S,\ A_S^S}$
$\boldsymbol{E_{A_S,\ B_S}}$	A_S	B_S	Reuse $E_{A_S,\ A_S}$

Table 4
The corresponding mean and standard deviation of image metrics reported for the corresponding violin plot are illustrated in Fig. 1(a). We highlight the entries between Australian (A_S) and Swiss (B_S) skin to illustrate the difference of the mean intensity values in the green and blue channels.

	Brightness	Contrast	RGB		HSV	
$\boldsymbol{A_S}$	0.75 ± 0.05	0.72 ± 0.07	**R**	0.81 ± 0.07	**H**	0.43 ± 0.03
			G	0.63 ± 0.09	**S**	0.69 ± 0.08
			B	0.78 ± 0.05	**V**	0.78 ± 0.10
$\boldsymbol{B_S}$	0.78 ± 0.05	0.77 ± 0.05	**R**	0.83 ± 0.04	**H**	0.53 ± 0.13
			G	0.72 ± 0.07	**S**	0.11 ± 0.04
			B	0.86 ± 0.03	**V**	0.90 ± 0.02
$\boldsymbol{B_L}$	0.65 ± 0.08	0.81 ± 0.40	**R**	0.86 ± 0.04	**H**	0.78 ± 0.03
			G	0.76 ± 0.05	**S**	0.15 ± 0.04
			B	0.88 ± 0.02	**V**	0.89 ± 0.02

Table 5
The p-value is the result of the Wilcoxon Rank-sum test of color distribution in terms of the brightness between (1) A_S and B_S; (2) B_L and B_S. We also measure the difference in the mean of Red, Green, and Blue intensity values. We highlight the entries to reflect that B_L and B_S are closer to that of B_S compared to A_S in terms of distance in the distribution of green and blue intensity values.

	p-value	Red	Green	Blue
$\boldsymbol{A_S}$ vs. $\boldsymbol{B_S}$	3×10^{-105}	0.02	0.09	0.08
$\boldsymbol{B_L}$ vs. $\boldsymbol{B_S}$	8×10^{-218}	0.03	**0.04**	**0.02**

identification of BCC and SCC-Invasive subtypes, DL-based approaches generally demonstrate clear superiority over non-DL-based approaches. Importantly, though, none of these works addresses the issue of distinguishing SCC-In Situ from SCC-Invasive or explicitly addresses the issue of decreased classifier performance due to domain shift when validated on data from external sites. Jiang et al. reported a drop in AUC (10% ∼ 20%) of their BCC recognition and segmentation model in both WSI and MOI cases when the model is applied to an external $\boldsymbol{D_V}$ (Jiang et al., 2019). To help address the discrepancy in cross-site performance of DL models, several domain adaptation approaches have been introduced (de Bel et al., 2019; Ganin and Lempitsky, 2014; Tellez et al., 2019). Some approaches perform domain adaptation by mapping the DL-extracted features in the source domain to the target domain (Ganin and Lempitsky, 2014). However, these approaches lack transparency, in the sense that it is difficult to visually assess and confirm the correctness of this mapping in the DL-feature space, and thus minimizes the opportunity for quality control (Ganin and Lempitsky, 2014). Other CycleGAN-based algorithms instead directly modify test images such that they appear similar to their training image counterparts (de Bel et al., 2019; Shaban et al., 2019; Zhu et al., 2017). These approaches can provide more transparency, as image modifications can be qualitatively (e.g., manual inspection) and quantitatively assessed. CycleGAN approaches have been shown to be sufficiently powerful in their ability to not only transfer the desired pre-analytical variables (e.g., stain) across domains, but also modify the overall image "style". For instance, a GAN has been shown to successfully "convert" arbitrary normal tissue images to corresponding realistic synthetic precancerous images (Wei et al., 2019). As previously alluded to, these approaches unfortunately may also facilitate *data leakage*. More critically, it remains difficult to quantify the extent and impact of such data leakage. Multi-Site Cross--Organ Calibrated Deep Learning (MuSCID) (illustrated in Fig. 2), the approach presented in this work, offers a series of unique advantages over previously proposed approaches and is the first approach that is evaluated in the context of multi-class NMSC-subtyping. Specific unique attributes of this manuscript are detailed below.

1 MuSCID builds on previous works in the NMSC-subtyping and domain adaptation fields (de Bel et al., 2019; Jiang et al., 2019; Kimeswenger et al., 2020; Marka et al., 2019; Shaban et al., 2019; Tellez et al., 2019). We extend our study in NMSC classification into a multi-site study to distinguish between BCC, SCC-In Situ, and SCC-Invasive subtypes.
2 We further demonstrate the impact of domain shift in multi-site NMSC-subtyping, and provide a cross-organ (skin/lung) calibration approach to mitigate this issue. This MuSClD approach differs from these previous works in that it is not (a) solely focused on the identification of BCC and SCC-Invasive but also focused on distinguishing SCC subtypes (i.e., SCC-In Situ versus SCC-Invasive), and (b) explicitly addresses potential data leakage endemic to many cross-site calibration based approaches.
3 By only modifying $\boldsymbol{D_T}$, the introduction of artifacts (e.g., blur, checkboard) into $\boldsymbol{D_V}$ is minimized, ensuring high-fidelity unaltered input data to the subtyping algorithm.
4 By calibrating in a cross-organ fashion, pre-analytical variables are exposed for learning while data leakage from the $\boldsymbol{D_V}$ to the $\boldsymbol{D_T}$ is largely avoided. This ensures a genuine independent evaluation of our NMSC-subtyping algorithm on the external testing site. Lastly, to the best of our knowledge, this work is the largest comprehensive study involving automated diagnosis and subtype classification of BCC and SCC from H&E images involving independent training and testing sites with over 400 unique patients.
5 We quantitatively illustrated the potential risk of data leakage during calibration. To our best knowledge, **there is no study pointing out such risk**, given the importance of calibration in multi-site studies.

3. Multi-Site cross-organ calibrated deep learning (MuSClD)

3.1. Notation

We use A to represent the data of the training site, and B to represent data from the testing site. Unless otherwise specified, we use a to represent individual images of site A, and b to represent images of site B. For the notation $\forall a \in A$ and $\forall b \in B$ defined above, we use the subscript τ to denote the type of tissue organ that is either skin (S) or lung (L) tissue ($\tau \in \{S,\ L\}$). For calibration outputs, we use the superscript μ to denote the type of tissue organ (S or L) that was employed as a template ($\mu \in \{S,\ L\}$). The relevant notation used in this work is illustrated in Table 1 below.

3.2. Algorithmic details of MuSCID

A CycleGAN-based calibration framework was chosen and can be thought of as representing the mapping between two domains: $A \rightarrow B$ and $B \rightarrow A$ (See Fig. 2a). Path $A \rightarrow B$ consists of a generator model G_{A2B} which accepts images produced from Site A and attempts to modify them such that they appear like images created from Site B (Goodfellow et al., 2014; Zhu et al., 2017). Path $A \rightarrow B$ also consists of a discriminator model D_B that attempts to determine for $\forall a \in A$, if $G_{A2B}(a)$ are distinguishable

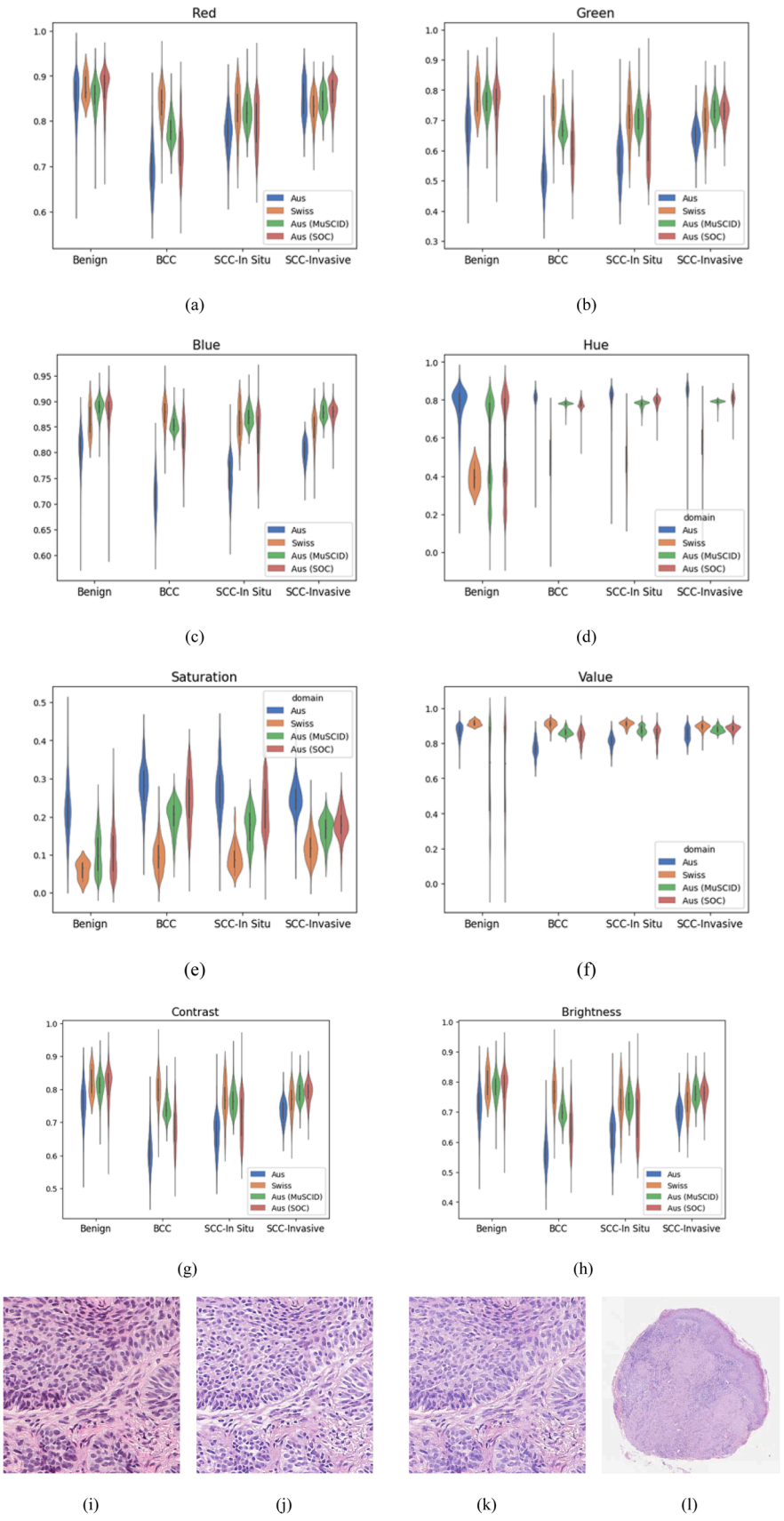


Fig. 3. Violin plots (a) to (h) illustrate the distribution of RGB intensity, the contrast, the brightness, and the HSV values of each class (benign/BCC/SCC-In Situ/SCC-Invasive) across original A_S (blue), B_S (orange), and A_S^L by MuSCID (green), and A_S^S by SOC (red), respectively. These statistics show that after the calibration, differences in contrast and color distribution are mitigated. The original Australian skin tissue (i), the A_S^L via MuSCID (j), the A_S^S via SOC (k), and the B_S show that after calibration (l), the Australian skin tissue more closely resembles the appearance of the Swiss skin tissue (bluish tinge). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Class-wise mean intensity of R/G/B channels of Australian skin tissue before (A_S) and after calibration (A_S^L and A_S^S), compared to Swiss skin tissue (B_S). We highlight the entries where a significant change in intensity values after calibration was observed, resulting in greater similarity between the Australian and Swiss site images.

		A_S	A_S^L	B_S	A_S^S
Benign	**R**	0.86 ± 0.05	0.86 ± 0.03	0.87 ± 0.03	0.87 ± 0.04
	G	0.69 ± 0.08	0.76 ± 0.05	0.77 ± 0.05	0.75 ± 0.07
	B	0.81 ± 0.04	0.89 ± 0.02	0.86 ± 0.02	0.88 ± 0.03
BCC	**R**	0.73 ± 0.06	0.78 ± 0.02	0.84 ± 0.04	0.74 ± 0.06
	G	0.57 ± 0.08	0.67 ± 0.04	0.74 ± 0.06	0.61 ± 0.07
	B	0.74 ± 0.04	0.85 ± 0.02	0.87 ± 0.02	0.83 ± 0.04
SCC-In Situ	**R**	0.79 ± 0.04	0.82 ± 0.03	0.82 ± 0.05	0.79 ± 0.05
	G	0.61 ± 0.08	0.71 ± 0.05	0.71 ± 0.07	0.63 ± 0.08
	B	0.77 ± 0.04	0.87 ± 0.02	0.86 ± 0.03	0.83 ± 0.04
SCC-Invasive	**R**	0.85 ± 0.04	0.84 ± 0.03	0.83 ± 0.03	0.86 ± 0.03
	G	0.65 ± 0.51	0.74 ± 0.04	0.70 ± 0.05	0.72 ± 0.04
	B	0.80 ± 0.02	0.88 ± 0.01	0.85 ± 0.03	0.87 ± 0.01

from real images $\forall b \in B$.

This process is aided by three loss functions (see **Equation 1**); the minimax GAN loss (L_{GAN}), the identity loss L_{Id}, and the cycle consistency loss (L_{Cyc}) as illustrated below.

$$\begin{aligned} \boldsymbol{L_{GAN}}(\boldsymbol{G_{A2B}}, \boldsymbol{D_B}, \boldsymbol{a}, \boldsymbol{b}) &= \mathbb{E}[\boldsymbol{logD_B}(\boldsymbol{b})] + \mathbb{E}[1 - \boldsymbol{logD_B}(\boldsymbol{G_{A2B}}(\boldsymbol{a}))] \\ L_{Id}(G_{A2B}, G_{B2A}) &= \mathbb{E}\left[||G_{A2B}(b) - b||_1\right] + \mathbb{E}\left[||G_{B2A}(a) - a||_1\right] \\ L_{Cyc}(G_{A2B}, G_{B2A}) &= \mathbb{E}\left[||G_{B2A}(G_{A2B}(a)) - a||_1\right] + \mathbb{E}\left[||G_{A2B}(G_{B2A}(b)) - b||_1\right] \end{aligned} \tag{1}$$

The notation $\mathbb{E}$ represents the expectation value. L_{GAN} represents the adversarial component balancing between the D_B and the G_{A2B}. It measures D_B's ability to recognize whether an image is from Site B or has been generated by G_{A2B}. Importantly, this loss is minimized when the Jensen-Shannon distance (Menéndez et al., 1997) between data from Site A matches those of Site B, after being modified by G_{A2B} (Sinn and Rawat, 2018). Intuitively, this implies that for $\forall a \in A$, the associated $G_{A2B}(a)$ is similar to images in B, suggesting that the calibration from A to B was successful. Terms L_{Cyc} and L_{Id} serve as regulators encouraging G_{A2B} to learn a model which produces images similar to those in B (Srivastava et al., 2017; Zhu et al., 2017).

As previously mentioned, to prevent G_{A2B} from learning task-specific knowledge from B (e.g., B_S), we choose an off-target organ (e.g., B_L) as a template to mitigate the potential for data leakage; as a result, the calibration strategy is "***cross-organ***". Previous works have shown however that GANs may affect the fidelity of tissue attributes in terms of morphology and cytology. For instance, GANs are known to introduce artifacts due to the loss of high-frequency content, and are also capable of learning histologic features other than pre-analytic variations (e.g.,

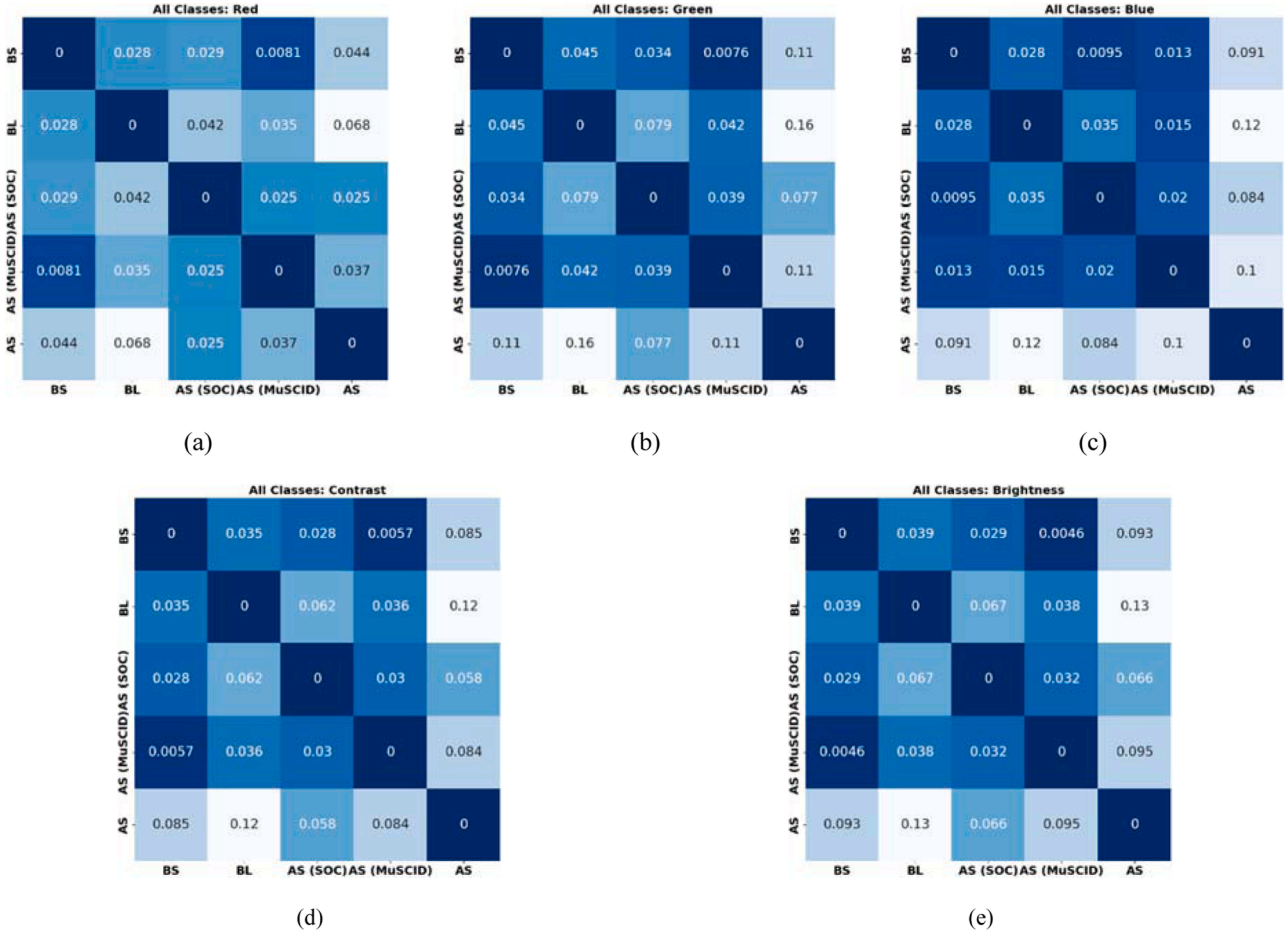


Fig. 4. Wasserstein distance of different color space measurements between datasets are calculated and plotted as heat maps: plots (a) to (e) correspond to red/green/blue intensity, contrast, and brightness measures respectively. From top left to bottom right, the 1st to 5th rows/columns of each heat map correspond to B_S, B_L, A_S^S, A_S^L, and A_S. Darker blue colors represent smaller distance values, and show that (c) calibrated images (A_S^L and A_S^S) are more similar to B_S and $B_L$compared to A_S in terms of the blue color distribution. Likewise, (b), (d), and (e) also suggest that after calibration, A_S^L and A_S^S are more similar to B_S compared to A_S in terms of contrast and green color intensity distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

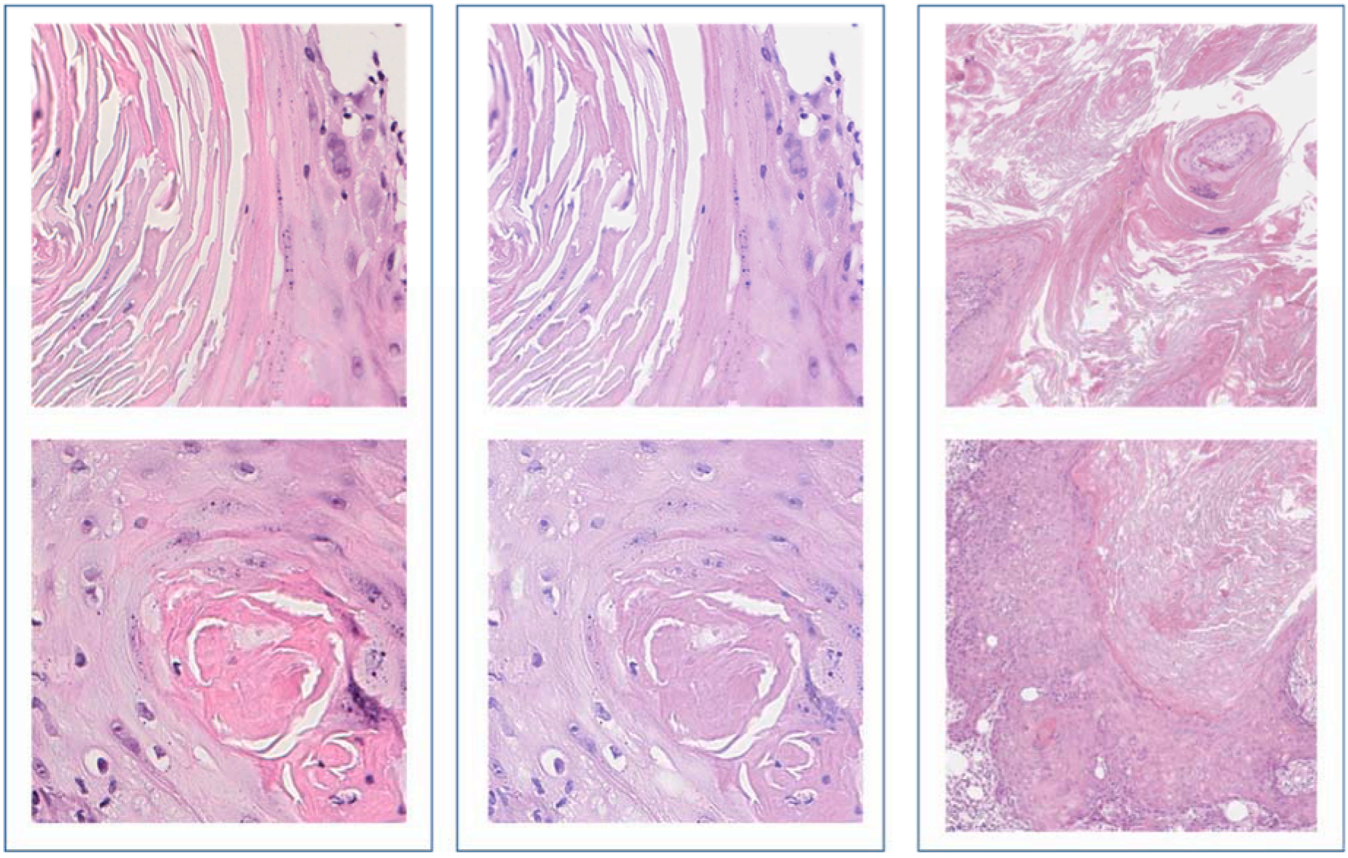


Fig. 5. Comparison of keratin (or region with keratinization) between (a) A_S (the 1st column), (b): A_S^L (the 2nd column), and (c) B_S (the 3rd column). All image patches showcased here are sampled from SCC-Invasive lesions. It may be observed that the (b) calibrated keratin is slightly bluer than the actual keratin in (c) Swiss data likely due to the lack of keratin in the lung template.

stain) from templates (Wei et al., 2019; Zhang et al., 2019). Consequently, an L1 reconstruction loss is added between the calibration input and outputs for both paths: a and $G_{A2B}(a)$, as well b and $G_{B2A}(b)$, $\forall a \in A;\ \forall b \in B$. We use the notation H_a, W_a, and C_a to denote the height, width, and channel depth of an arbitrary input image a, respectively, the L1 reconstruction loss, which can then be defined as follows:

$$\widetilde{L} = \frac{\mathbb{E}\left[||G_{A2B}(a) - a||_1\right]}{H_a \times W_a \times C_a} \tag{2}$$

Here $\widetilde{L}$ helps to quantitatively measure and monitor pixel-wise changes between a and $G_{A2B}(a)$, during and after training of MuSCID. Pixel values during the computation of $\widetilde{L}$ in Eq. (2) are normalized to be within the range [0,1]. The idea of a reconstruction loss is to help preserve the stylistic and spatial attributes in the tissue of $a \in A$ after calibration. This approach was previously described (Johnson et al., 2016), where an L1 loss was chosen for its empirically demonstrated superior performance in obtaining high-fidelity output images (Zhao et al., 2015).

4. Experimental design

4.1. Dataset description

A summary of the cohorts employed in this study is illustrated in Table 2. The datasets comprised H&E WSI collections obtained from two international sites: (a) Cohort from site A ($n = 85$) patients used for training from Southern Sun Pathology, Australia, and (b) Cohort from site B ($n = 352$) patients from Kantonsspital Aarau, Switzerland used for testing. To model the pre-analytic properties anticipated in cohort B, four WSIs of lung specimens (B_L) were employed from the B to calibrate skin slides from A (A_S). B_L were prepared in a similar manner (e.g., stainer, tissue thickness, and slide scanner) as B_S, while coming from patients not part of the NMSC cohort. All data in B was collected at 40x magnification and down-sampled to 20x in order to approximate the resolution of the images originating from A.

4.2. NMSC-subtyping implementation details

A ResNext50_32 × 4D architecture was selected for the NMSC-subtyping DL classifier. This classifier was chosen given its previously demonstrated high predictive performance with comparatively few parameters compared to other DL architectures (Xie et al., 2017). Image patches of size 512 × 512 were extracted at 20x from regions of tumor annotated by human dermatopathologists in slides from site A. As is common practice, ***stain augmentation*** was employed during training of all DL networks (details in the Section S1). Patient-level predictions were generated by aggregating patch-level predictions throughout the WSIs (Hou et al., 2016). Briefly, each patch was classified by the NMSC subtype classifier into one of four target classes: benign, BCC, SCC-In Situ, and SCC-Invasive. The ResNext50_32 × 4d classifier outputs the class prediction scores for each image patch. For each of the three cancer classes, a 32-bin histogram is created to aggregate the patch-level raw class output values (see Section S1). Each of the three histograms corresponding to the three cancer classes was then normalized and concatenated, resulting in a 1 × 96 vector signature for each patient.

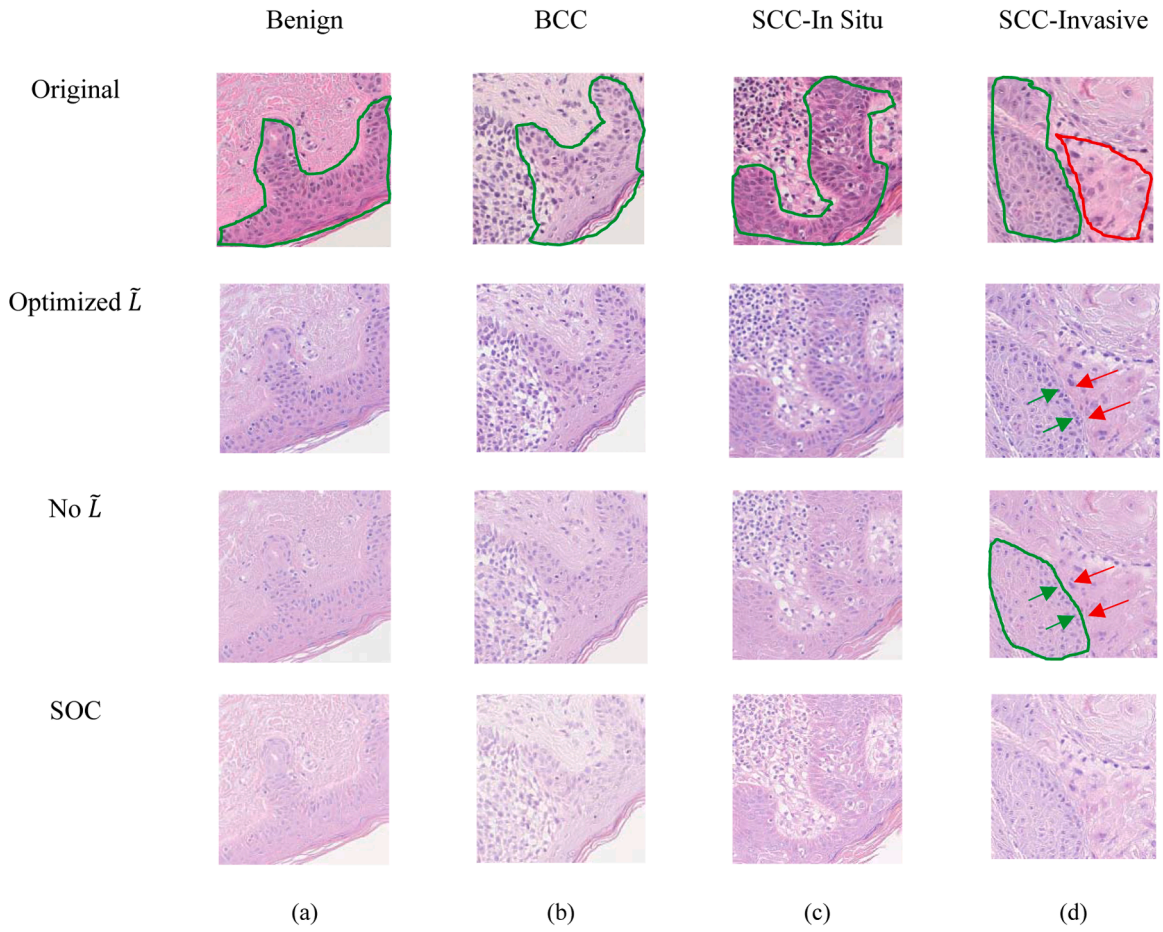


Fig. 6. Comparison of the calibration output with and without optimizing $\tilde{L}$ where columns (a) to (d) demonstrate benign, BCC, SCC-In Situ, and SCC-Invasive exemplars respectively. The first row illustrates the original A_S. The second and third rows respectively show the calibrated output with and without $\tilde{L}$ being optimized. For reference, the fourth row is the corresponding output of SOC. It may be observed in (a) to (d) that the boundary of the epidermis is more pronounced when $\tilde{L}$ is optimized. More specifically in (d), without $\tilde{L}$, the color of the thickened epidermis within the green annotated region (3rd row) mostly resembles that of the dermis (the boundary between epidermis and dermis marked by red/green arrows), as opposed to its counterpart with optimized $\tilde{L}$ (boundary marked by red/green arrows). Furthermore, without optimizing the $\tilde{L}$, similar color artifacts also occur in the SCC-In Situ showcase produced by SOC (the 4th row and the 3rd column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This vector was employed as the input for training a fully connected neural network for predicting the final patient-level output.

MuSCID was trained until the reconstruction loss $\tilde{L}$ between a and $G_{A2B}(a)$ converged, and $G_{A2B}(a)$ qualitatively resembled the pre-analytic appearance (e.g., stain) of B_L. The detailed network configurations and the hyperparameters are listed in Section S1.

4.3. Experiment 1: Evaluation of MuSCID in terms of consistency of skin tissue components, color distribution, and potential data leakage

The main premise behind Experiment 1 is that differences in image presentation between NMSC images from A and B can be ameliorated with MuSClD by using B_L images from the Swiss site while not imbuing it with significant histologic variations. A variety of evaluation strategies were employed for Experiment 1 as described below.

4.3.1. Inter- and intra-site color distribution of skin slide

Image metrics capturing the (a) brightness, (b) root mean square (RMS) contrast, and (c) mean intensity of each of the red/green/blue (RGB) and hue/saturation/value (HSV) channels are computed from the skin and lung slides of both sites A_S and B_S to examine the relative inter-site variability. Swiss lung images, B_L, are also analyzed in a similar fashion to determine their suitability for serving as template surrogates for B_S. Violin plots were employed to visualize the image metric distributions. Besides the distribution distance, a Wilcoxon rank-sum test was also used to determine if image metric distributions between A_S and B_S are statistically significant.

4.3.2. Inter-site and class-wise color distribution between calibrated Australian skin images and Swiss images

MuSClD was employed to calibrate Australian NMSC images with B_L (Swiss lung). If MuSClD is successful at compensating for pre-analytical variance between A and B, the distance between image metrics from A_S^L and B_S should be reduced. To demonstrate this reduction for each subtype, two image metrics for all four subtypes (benign/BCC/SCC-In Situ/SCC-Invasive) are computed between A_S^L and B_S. First, Wasserstein distance, a commonly invoked metric for evaluation of domain shift (de Bel et al., 2019; Stacke et al., 2019; You et al., 2020b), was employed to compare RGB and contrast values, wherein smaller Wasserstein distances indicate greater similarity. Additionally, A Wilcoxon rank-sum test was used to determine if the image metrics of A_S^L are statistically different from A_S. A reduction in Wasserstein distance between A_S^L and B_S would suggest that calibration was performed successfully and the previously identified domain shift had been ameliorated. It should be noted that, since Wilcoxon rank-sum test also considered other factors of distributions (e.g., distribution shape) than their distances, a statistically

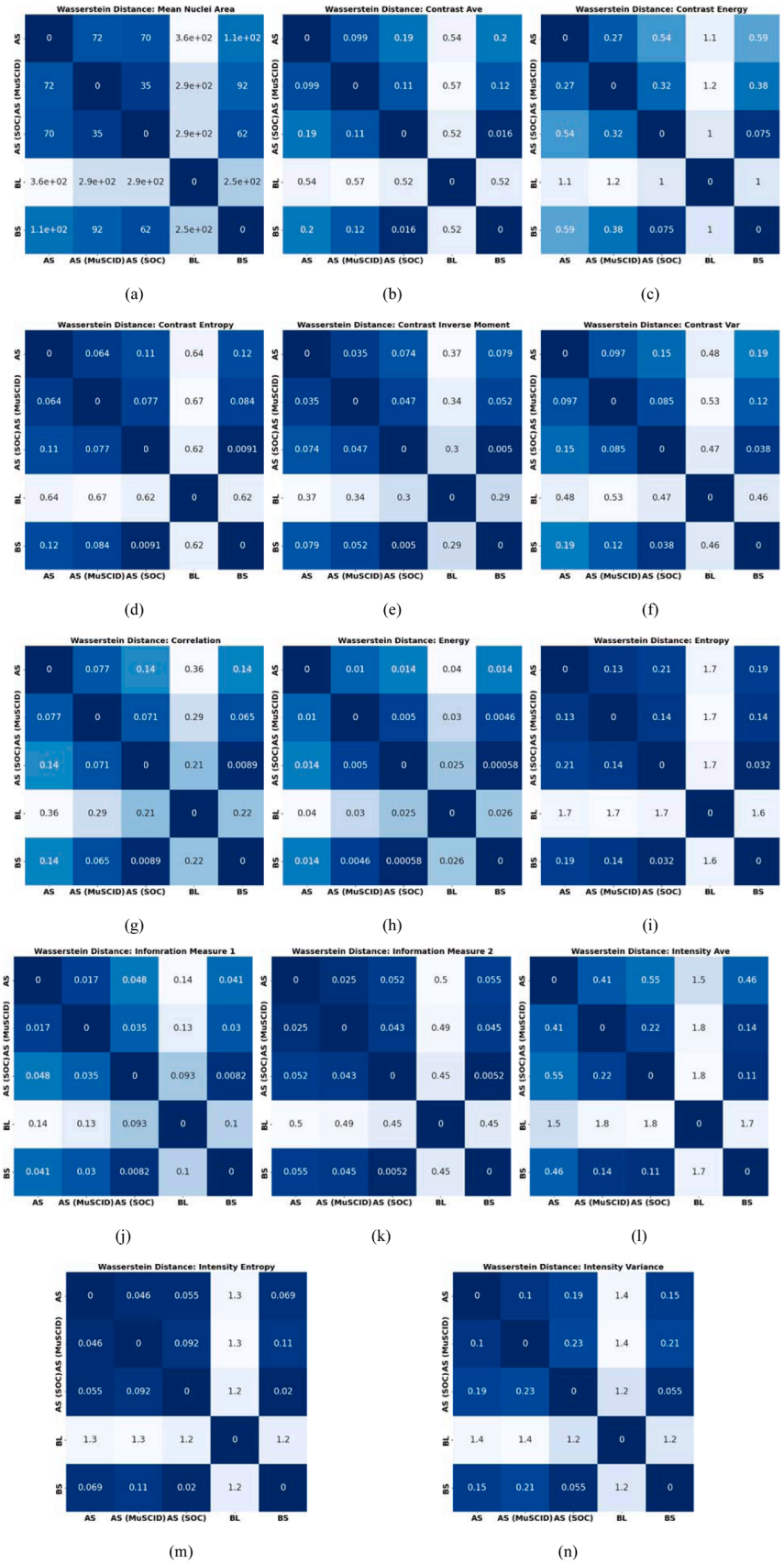


Fig. 7. Wasserstein distance of different morphological features between datasets are calculated and plotted as heat maps. From top left to bottom right, the 1st to 5th rows/columns of each heat map correspond to A_S, A_S^L, A_S^S, B_L, and B_S. Darker blue colors represent smaller distance values, i.e. higher similarity. Plot (a) represents the mean nuclei area per patch extracted in each dataset. Plots (b) to (n) correspond to 13 nuclei-wise Haralick texture features extracted from tiles in all datasets. It may be observed that, all skin tissue datasets (A_S, A_S^L, A_S^S, and B_S) are relatively similar to each other compared to B_L, in terms of the morphological feature. This in turn suggests that the calibrated output reasonably resembles the skin-specific morphological information. Meanwhile, shown in all plots, the similarity between A_S^S to B_S are smaller compared to between A_S^L and B_S, demonstrating that employing skin images as the template may potentially incur the risk of leaking morphological information from B_S into A_S^S. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
(a) Patch-level and (b) Patient-level one-vs.-rest multiclass AUC of comparison strategies for the training and internal validation. The AUC of $E_{A_S,\ A_S}$ is close to $E_{A_S^L,\ A_S^L}$, indicating only limited error was introduced into $\boldsymbol{D_T}$ by MuSCID, which may otherwise potentially degrade the model performance.

Patch-level AUCs	Benign	BCC	SCC-In Situ	SCC-Invasive
$E_{A_S,\ A_S}$	0.99	0.99	0.97	0.99
$E_{A_S^L,\ A_S^L}$	**0.98**	**0.99**	**0.94**	**0.98**
$E_{A_S,\ A_S^L}$	0.98	0.91	0.76	0.97
$E_{A_S^L,\ A_S}$	0.98	0.96	0.80	0.95
$E_{A_S^S,\ A_S^S}$	0.98	0.99	0.92	0.86
(a)				
Patient-level AUCs	Benign	BCC	SCC-In Situ	SCC-Invasive
$E_{A_S,\ A_S}$	0.97	0.98	0.92	0.97
$E_{A_S^L,\ A_S^L}$	**0.96**	**0.97**	**0.91**	**0.98**
$E_{A_S,\ A_S^L}$	0.99	0.94	0.78	0.91
$E_{A_S^L,\ A_S}$	0.97	0.96	0.88	0.99
$E_{A_S^S,\ A_S^S}$	0.90	0.95	0.85	0.94
(b)				

Table 8
(a) Held-out test of the ResNext50_32 × 4D NMSC-subtyping model with and without calibrating the $\boldsymbol{D_T}$ using lung tissue templates from the testing site. (b) The held-out test result for the SOC counterparts and the corresponding Delong test p-value against the MuSCID. We highlighted the statistically significant p-value (SCC-Invasive), suggesting that the performance of NMSC-subtyping between SOC and MuSCID is significantly different only in SCC-Invasive cases, with a 3% difference in AUC scores. The performance between SOC and MuSCID in regards to the AUC of NMSC-subtyping is similar in benign, BCC, and SCC-In Situ.

	Benign (p = 0.47)	BCC (p = 0.01)	SCC-In Situ (p = 0.15)	SCC-Invasive (p = 1e-5)
$E_{A_S^L,\ B_S}$	0.92	**0.92**	**0.87**	**0.92**
$E_{A_S,\ B_S}$	**0.96**	0.87	0.73	0.82
(a)				
$E_{A_S^S,\ B_S}$	AUC	p-value (vs. $E_{A_S^L,\ B_S}$)		
Benign	0.98	0.40		
BCC	0.95	0.052		
SCC-In Situ	0.85	0.84		
SCC-Invasive	**0.95**	**0.03**		
(b)				

significant p-value between A_S^L and B_S do not suggest that the calibration is unsuccessful (S2.7).

4.3.3. Consistency of skin histologic components pre- and post-calibration

To evaluate the benefit of including the reconstruction loss $\widetilde{L}$ in MuSCID, we both qualitatively and quantitatively compared $\widetilde{L}$ in the inference stage with and without optimizing $\widetilde{L}$ during the training stage. We hypothesized that the inclusion of $\widetilde{L}$ during the model optimization process would encourage the retention of consistency in the histologic appearance of the skin while limiting the introduction of lung morphology during the calibration of A_S to A_S^L.

To visually and qualitatively assess the impact of calibration on A_S^L, 512 4096 × 4096 regions of interest (ROIs) were inspected by a dermatopathologist. These 512 ROIs consisted of: 51 benign, 336 BCC, 70 SCC-In Situ, and 55 SCC-Invasive cases. During the inspection, the dermatopathologist was informed of the disease type of the ROIs, and requested to report any observed skin histology alterations or calibration artifacts.

For further comparison, skin templates B_S were employed for calibration to produce A_S^S. Discrepancies between A_S^L and A_S^S were visually evaluated in terms of tissue texture and color space qualities to better understand the potential consequences of data leakage.

4.3.4. Similarity measurement of morphological feature distribution pre- and post-Calibration

To the best of our best knowledge, a direct method to quantitatively assess data leakage during calibration does not exist. As a surrogate measure, Haralick texture features and mean nuclei area between A_S^L and A_S^S are compared. This allows us to estimate whether calibration introduces different morphological information into A_S based on the template organ of choice.

The publicly available HoverNet (Graham et al., 2018) model was used to perform nuclei segmentation. This model was selected as it was trained on over 200k nuclei from multiple data sources, and showed consistently robust model performance during validation. Here we leverage that generalizable performance across our datasets (A_S, A_S^L, A_S^S, B_S, and B_L) to ensure that nuclei-level features were mostly captured in comparable nuclear regions.

Thirteen nuclear-specific Haralick texture features (Haralick et al., 1973) were extracted within the segmented nuclei regions, along with the mean nuclei area. Similar to Section 4.3.2, the Wasserstein distance between each of the feature distributions was computed. Additionally, to measure whether these morphological features are significantly different from each other, the Wilcoxon Rank-sum p-value of these feature distributions between datasets (A_S, A_S^L, A_S^S, B_S, and B_L) was computed, with a p-value of 0.05 set to determine statistical significance.

4.4. Experiment 2: Comparative evaluation of M_c and M_n in terms of AUC for NMSC diagnosis and classification

We evaluate the NMSC-subtyping models trained with A_S and A_S^L, respectively on held-out $\boldsymbol{D_V}$ sampled from B_S and compare their subtyping performance in terms of the area under the curve (AUC) of the corresponding receiver operating characteristic (ROC) curves produced by thresholding the output prediction score of NMSC-subtyping models. The one –vs. rest multi-class AUC (Rifkin and Klautau, 2004) is employed as the metric for the prediction. During the held-out testing phase, all patches from the tissue region without artifacts (blurry, pen-marker) are fed into the NMSC-subtyping model. Quality control of patches was assessed using HistoQC (Janowczyk et al., 2019), a WSI quality control tool (e.g., blurry and tissue folding identification), along with the manual inspection. The Delong test was used to assess whether the improvement between ROC curves is statistically significant (DeLong et al., 1988). An additional set of sub-experiments, examining combinations of $\boldsymbol{D_T}$ and internal $\boldsymbol{D_V}$ composed of A_S and A_S^L were investigated (see Table 3).

4.4.1. Internal validation and evaluation of D_v

$\boldsymbol{D_T}$ from A was randomly split at the patient level into $\boldsymbol{D_T}$ and internal $\boldsymbol{D_V}$ using a ratio of 7:3, such that the distribution of each subtype is preserved. In the first four experiments in Table 3, models were trained and validated on the images coming from the same site, A. The rationale of these experiments is that if $E_{A_S^L,\ A_S^L}$ has similar performance metrics to that of $E_{A_S,\ A_S}$ then any potential image artifacts introduced by the calibration process minimally affected subtype prediction performance. On the other hand, a difference between performances in $E_{A_S^L,\ A_S^L}$ vs. $E_{A_S,\ A_S^L}$ and $E_{A_S,\ A_S}$ vs. $E_{A_S^L,\ A_S}$, would suggest that domain shift between A_S and A_S^L degrades the ability of models to generalize. The impact of domain shift between A and B, as well as the benefit of calibration, is evaluated in the held-out tests $E_{A_S^L,\ B_S}$ and $E_{A_S,\ B_S}$. To better understand the impact of potential data leakage, B_S was employed as a template to calibrate A_S to produce A_S^S, this strategy has been previously employed in the past (de Bel et al., 2019; Shaban et al., 2019). An NMSC-subtyping model was subsequently trained with A^S. Similar AUC values between

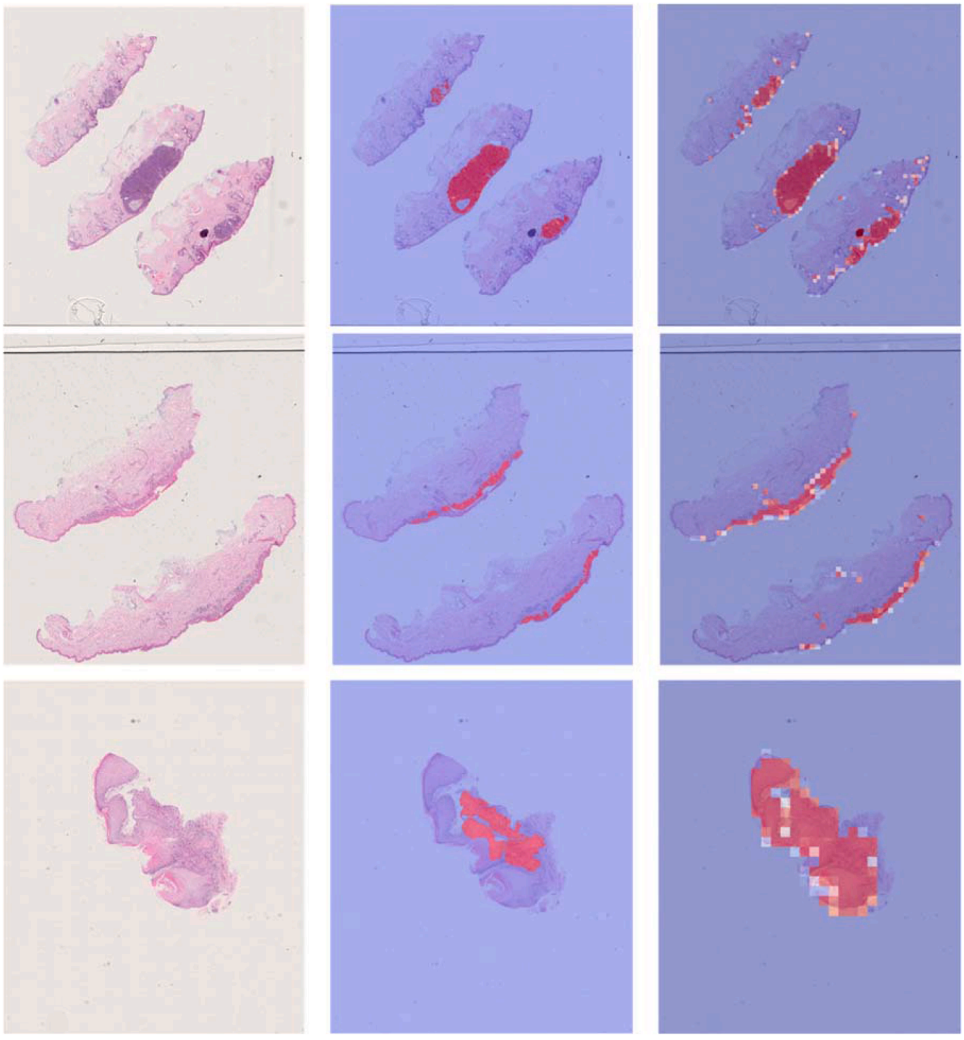


Fig. 8. Example of the WSI-level saliency maps produced using the NMSC-subtyping model on A_S^L, with the rows from top to bottom illustrating examples of BCC, SCC-In Situ, and SCC-Invasive cases respectively. Column (a) is the original WSI, (b) highlights the ground truth annotated by the dermatopathologist, and (c) shows the identified regions of BCC, SCC-In Situ, and SCC-Invasive respectively. The red color illustrates those regions identified with a higher probability of cancer, and the blue represents those regions identified by the classifier as being more likely to be benign. It may be observed when comparing (b) and (c) that a large majority of the cancerous regions are successfully identified by the NMSC-subtyping network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cross-site on- and off-target organ tests (i.e., $E_{A_S^S,\ B_S}$ and $E_{A_S^L,\ B_S}$) would suggest that MuSCID effectively mitigates the domain shift in terms of the NMSC-subtyping performance, while also minimizing the risk of data leakage. The difference between the AUC values $E_{A_S^S,\ B_S}$ and $E_{A_S^L,\ B_S}$ was again evaluated by the Delong test.

4.4.2. Visual interpretation of NMSC-subtyping model

DL models are often considered "black-box" since there is a lack of interpretability or understanding of what DL models learn in order to succeed. To obtain a WSI-level view of the DL's capability of tumor localization, a heat-map of patch prediction scores is generated and overlaid on the original image. Grad-CAM is used to visualize regions in the patches that are most informative to the model when making a prediction (Selvaraju et al., 2016) (see S2.3 for more details). If these highlighted regions contain relevant information for identifying NMSC subtypes, it suggests that our DL model has successfully learned clinically discriminating features. Lastly, 2D t-distributed Stochastic Neighbor Embedding (t-SNE) plots of the high dimensional feature space created by the NMSC-subtyping model are shown (Arora et al., 2018) (see S2.1). The t-SNE shows whether the clustering of features is stratified by the disease type, and potentially identifies difficult cases falling on the boundary between the subtype clusters.

5. Results and discussion

5.1. Experiment 1: Evaluation of MuSCID in terms of consistency of skin tissue components, color distribution, and potential data leakage

5.1.1. Inter- and intra-site color distribution of skin slide

The impact of domain shift across the Australian (A) and Swiss (B) sites is evident, observable by the inconsistency between the intensities of the green and blue channels in Table 4. Fig. 1(a) illustrates the necessity for calibration to mitigate the impact of such domain shift for the multi-site generalizability of NMSC-subtyping between A and B. The violin plot in Fig. 1(a) also shows that the intra-site color distribution difference between B_S and B_L is less severe than that of the inter-site color distribution difference between A_S and B_S, especially in terms of green and blue channel intensity statistics. There is also a large difference in the saturation channel intensity statistics of the HSV representation between A_S and B_S. Moreover, the Wilcoxon rank-sum test in Table 5 also illustrates that the domain shift in terms of the difference of color distribution between A_S and B_S is significant ($p = 3 \times 10^{-105}$). Interestingly, Table 5 shows that B_L is much closer to B_S than A_S in terms of the green (55% closer) and blue channels (75% closer), suggesting that B_L may be suitable as a calibration template for B_S.

5.1.2. Inter-site and class-wise color distribution between calibrated Australian skin images and swiss images

Fig. 3, Table 6, and Fig. 4 show that post-calibration, the distribution of color metrics in A_S^L and A_S^S are closer to B_S and B_L compared to A_S, especially in terms of the green (as much as 69% and 93% closer) and blue channels (85% and 89% closer). This suggests that the calibration indeed dampened the domain shift in both color channels. Fig. 4 and violin plots in Fig. 3 suggest that A_S^L and A_S^S are also similar in terms of color metrics, indicating MuSCID might be a suitable replacement for SOC.

Keratinization is a potential explanation for the greater green and blue intensities in the Australian versus the Swiss SCC-Invasive lesions (see Table 6). Keratinization and keratin pearls are commonly observed in well-differentiated SCC-Invasive, and have a pink/red appearance in both A_S and B_S (Fig. 5(a) and (b)) (Brown et al., 1979; Samarasinghe and Madan, 2012; Yanofsky et al., 2011). However, keratin only exists in lung tissue under abnormal circumstances (e.g., lung squamous cell carcinoma), and no keratin was observed in B_L. Subsequently, due to the absence of keratin in B_L templates, the calibration of keratin regions in A_S is likely affected, resulting in the keratin in SCC-Invasive of A_S^L appearing bluer than the keratin in SCC-Invasive of B_S, and thus potentially impacting the blue channel metrics (see Fig. 5 and Table 6). This suggests that a good candidate for off-target template organs is one that resembles the on-target organ, in terms of tissue type (e.g., epithelial or not), stain, and composition of tissue histologic structures (e.g., whether keratinization takes place). The keratinizing regions in A_S^S more closely resemble the color of keratin pearls compared to A_S^L (see Section S2.6), possibly due to the absence of keratinizing regions in B_L during calibration of A_S^L.

5.1.3. Consistency of skin histologic components pre- and post-calibration

When the objective $\widetilde{L}$ is minimized during training, the final averaged value of $\widetilde{L}$ across all pairs of images in A_S^L and A_S was 0.076 ± 0.021, versus 0.085 ± 0.025 when it is not minimized. This 11% quantitative improvement appears to coincide with an improvement in the consistency of skin histology components (see Fig. 6). For example, when not minimizing $\widetilde{L}$, epidermis regions presented in A_S typically have darker colors compared to the dermis regions. With the optimization of $\widetilde{L}$, the epidermis (circled in green) in calibrated outputs of Fig. 6(a)-(c) are of similar color to the remaining dermis regions (circled in red). Fig. 6(d) shows that, without $\widetilde{L}$, epidermis region color (green arrows) more closely resembles that of the dermis region (red arrows). During the inspection of the ROIs by the dermatopathologist, skin histologic features were appropriately preserved, with no lung-specific histology being introduced into the calibrated skin tissue. In additional calibration examples, subtle changes in tissue texture details (e.g., nuclei textures) between MuSCID and SOC are highlighted (see S2.6 and Fig. S7). Importantly, these changes were inspected by the dermatopathologist, and nothing unreasonable was identified from a biological standpoint, suggesting MuSCID is appropriate for color calibration.

5.1.4. Similarity measurement of morphological feature distribution pre- and post-calibration

As shown by the Wilcoxon Rank-sum test p-values (see Fig. S8), morphological feature distributions between A_S, A_S^S, A_S^L, B_L, and B_S are all significantly different from each other. Ideally, however, morphological differences between A_S, A_S^S, and A_S^L should not be statistically significant, since they contain the same images exposed to different sets of pre-analytic variations. This indicates CycleGAN may transmit different morphological features from templates images (B_L or B_S) to A_S. To evaluate if such transmission is related to data leakage, more fine-grained comparisons were performed.

First, all Wasserstein distance calculations between skin tissue datasets (A_S, A_S^S, A_S^L, and B_S) are notably smaller than compared to B_L, an expected finding given that skin and lung cells present differently. For instance, in terms of the mean nuclei area in Fig. 7(a), the Wasserstein distance from A_S^L to A_S is 71.686, which is about 75% smaller as compared to B_L (291.082). This quantitatively supports the dermatopathologist's visual findings with respect to the similarity of uncalibrated and calibrated skin tissue images (see Section 5.1.3), suggesting that both SOC and MusCID preserve morphological patterns that are specific to the skin.

Further, these results support the notion that SOC may transmit skin-specific morphological information from B_S into A_S^S. For example, Fig. 7 (a) to (n), shows the distance between A_S^S and B_S being smaller than that between A_S^L to B_S, with a mean nuclei size of A_S^S 31% closer to B_S (distance 62.094) compared to from A_S^L to B_S (distance 91.545). This smaller distance suggests the presence of data leakage in SOC from B_S templates.

Taken together, these findings suggest that potential **data leakage** in SOC use cases may not be adequately ruled out, motivating the need for off-target normalization processes like MusCID. Examining Wasserstein distances between B_L and B_S, reveals **the least amount of shared morphological information**. This further provides evidence that the risk of data leakage in MuSCID is minimized due to organ-specific information being unavailable for transmission from the onset. Moreover, A_S^L reasonably preserves skin morphological features, in turn suggesting that MuSCID is an appropriate surrogate for SOC.

5.2. Experiment 2: Comparative evaluation of M_c and M_n in terms of AUC for NMSC diagnosis and classification

5.2.1. Internal validation and evaluation of D_V

The similarity in AUC between $E_{A_S^L,\ A_S^L}$ and $E_{A_S,\ A_S}$ suggests minimal error is introduced into the $\boldsymbol{D_T}$ during the calibration process (see Table 7). On the other hand, differences in AUC were witnessed when comparing $E_{A_S,\ A_S^L}$ to $E_{A_S,\ A_S}$, suggesting that the domain shift between A_S and A_S^L in $\boldsymbol{D_V}$ is sufficient to degrade the performance of the NMSC-subtyping model. Conversely, when the $\boldsymbol{M_C}$was employed on A_S ($E_{A_S^L,\ A_S}$ vs $E_{A_S^L,\ A_S^L}$), diminished performance was again observed, as evidenced by patch-level AUCs. Taken together, these results support our hypothesis that domain shift degrades the performance of DL-based NMSC-subtyping models. Corresponding ROC curves of the AUC values in Table 7 are illustrated in Section S2.9.

Next, to demonstrate that MuSCID aids in mitigating the domain shift between Australian and Swiss skin tissue images, AUC values between $E_{A_S^L,\ B_S}$ and $E_{A_S,\ B_S}$ were compared and found to be statistically significant for the subtyping of BCC and SCC-Invasive classification problems (see Table 8). While an improvement in AUC of almost 14% was observed after calibration, the improvement was not found to be statistically significant for SCC-In Situ cases ($p = 0.15$), likely due to the limited sample size ($n = 9$ for benign and $n = 12$ for SCC-In Situ). Comparing $E_{A_S^S,\ B_S}$ to $E_{A_S^L,\ B_S}$ resulted in a statistically significant improvement only for the SCC-Invasive cases, with a 3% larger AUC value. This appears to suggest that MuSCID and SOC are similar in their performance of mitigating domain shift. Hence, MuSCID appears to be a suitable replacement for the more commonly employed SOC approaches, with the added benefit of further minimizing the risk of data leakage.

5.2.2. Visual interpretation of NMSC-subtyping model

In addition to quantitative evaluation, a qualitative heat-map visualization of regions identified as BCC/SCC-In Situ/SCC-Invasive was also provided (see Fig. 8). The NMSC-subtyping model appears to adequately capture the cancerous regions. False positives, especially in SCC cases, do exist in certain regions due to the sectioning of the specimen and the lack of context (see Section S3).

6. Conclusion

In this work, we presented Multi-Site Cross-Organ Calibrated Deep Learning (MuSCID), a new approach to mitigate the domain shift between multiple sites and applied it to non-melanoma skin cancer (NMSC) subtyping use case. MuSCID appears to effectively mitigate domain shift across histopathological images from different sites, aiding in the quest for increased generalizability of DL-based computational pathology approaches. Specifically in this work, MuSCID was found to aid in the identification of cases of benign, basal cell carcinoma (BCC), in situ squamous cell carcinoma (SCC-In Situ), and invasive squamous cell carcinoma (SCC-Invasive). We evaluated the performance of MuSCID by (1) assessing changes in color-based image metrics pre- and post-calibration of training images; (2) examining the improved generalizability of the NMSC-subtyping model across sites afforded by calibration; (3) comparing the color distribution and held-out test AUC score of MuSCID to the SOC. We show that cross-organ calibration can aid in mitigating domain shift while mitigating the risk of data leakage. To the best of our knowledge, this multi-site NMSC-subtyping study is the largest to date, involving over 400 patients ($n = 437$) curated from two international institutions.

Our study demonstrates that MuSCID performs comparably to more common SOC approaches both qualitatively and quantitatively, and thus may be able to act as a SOC replacement that minimizes data leakage risk. Notably, our approach only modifies $\boldsymbol{D_T}$ in the calibration procedure, mitigating the likelihood of artifact introduction into $\boldsymbol{D_V}$. Interestingly, we show that employing cross-organ calibration does not affect the underlying histologic fidelity of the $\boldsymbol{D_T}$. Specifically, this study shows that lung tissue can be employed for the calibration of skin tissue, and further aids in mitigating the decrease in performance due to the domain shift of an NMSC-subtyping model.

Previous studies regarding the automated identification of NMSC mostly focused on BCC and/or invasive SCC cases only (Jiang et al., 2019; Kimeswenger et al., 2020; Marka et al., 2019), and demonstrated that the performance of NMSC-subtyping suffers from domain shift. Here when employing MuSCID, the performance of a DL-based NMSC-subtyping was on par with previous intra-site studies, while also mitigating the effects of domain shift in cross-site evaluation. Moreover, this study expands prior NMSC classification problems (Marka et al., 2019) by the inclusion of SCC-In Situ disease. Our approach does not increase or decrease data storage overhead. While our approach performed well when identifying most of the different skin cancer subtypes, it was less successful in identifying SCC-In Situ disease; potential future directions to address this issue might involve characterizing the spatial context of the tumor bed. An additional direction for exploration in regard to MuSCID might include an evaluation of which organs are most apt to serve as templates for color calibration. For instance, we observed that organ-specific keratinization resulted in a less accurate coloring of keratin regions in SCC-Invasive cases. While this did not appear to affect the overall performance of the SCC-Invasive detector, the impact of template choice on the calibration quality clearly needs further and more rigorous study. While the availability of cross-organ WSIs is a strict requirement of our framework, many pathology laboratories and research facilities focus on multiple tissue/disease types. So while certain isolated specialized centers may not be able to take advantage of the improved experimental workflows afforded by MuSCID, we believe generalized facilities like hospitals and pathology labs will organically stand to benefit.

Overall, our study shows that MuSCID successfully mitigates the domain shift between two sites of skin data by employing off-target organ calibration. Data calibrated with MuSCID showed improved subsequent cross-site NMSC-subtyping performance, while minimizing the potential risk of data leakage.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute under award numbers 1U24CA199374-01, R01CA249992-01A1, R01CA202752-01A1, R01CA208236-01A1, R01CA216579-01A1, R01CA220581-01A1, 1U01CA239055-01, 1U01CA248226-01, 1U54CA254566-01, National Heart, Lung and Blood Institute 1R01HL15127701A1, National Institute for Biomedical Imaging and Bioengineering 1R43EB028736-01, National Center for Research Resources under award number 1 C06 RR12463-01, VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service, the Office of the Assistant Secretary of Defense for Health Affairs, through the Breast Cancer Research Program (W81XWH-19-1-0668), the Prostate Cancer Research Program (W81XWH-15-1-0558, W81XWH-20-1-0851), the Lung Cancer Research Program (W81XWH-18-1-0440, W81XWH-20-1-0595), the Peer Reviewed Cancer Research Program (W81XWH-18-1-0404), the Kidney Precision Medicine Project (KPMP) Glue Grant, the Ohio Third Frontier Technology Validation Fund, the Clinical and Translational Science Collaborative of Cleveland (UL1TR0002548) from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH roadmap for Medical Research, and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102702.

References

Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., Shi, W., 2017. Checkerboard artifact free sub-pixel convolution: a note on sub-pixel convolution, resize convolution and convolution resize.

Arora, S., Hu, W., Kothari, P.K., 2018. An analysis of the t-SNE algorithm for data visualization.

Bel, T.de, Bokhorst, J.-.M., Laak, J.van der, Litjens, G., 2021. Residual cyclegan for robust domain transformation of histopathological tissue slides. Med. Image Anal. 70, 102004 https://doi.org/10.1016/j.media.2021.102004.

Brown, C.L., Klaber, M.R., Robertson, M.G., 1979. Rapid cytological diagnosis of basal cell carcinoma of the skin. J. Clin. Pathol. 32, 361–367.

Chen, Y., Janowczyk, A., Madabhushi, A., 2020. Quantitative assessment of the effects of compression on deep learning in digital pathology image analysis. JCO Clin. Cancer Inform. 4, 221–233.

Chiavegatto Filho, A., Batista, A.F.M., Dos Santos, H.G., 2021. Data leakage in health outcomes prediction with machine learning. comment on "Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J. Med. Internet Res. 23, e10969.

de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G., et al., 2019. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: Cardoso, M.J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., et al. (Eds.), Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research. PMLR, pp. 151–163. London, United Kingdom.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.
Dodge, S., Karam, L., 2016. Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). https://doi.org/10.1109/qomex.2016.7498955.
Dong, Q., 2022. Leakage prediction in machine learning models when using data from sports wearable sensors. Comput. Intell. Neurosci. 2022, 5314671.
Elder, D.E., Massi, D., Scolyer, R.A., 2018. WHO Classification of Skin Tumours. International Agency For Research On Cancer.
Foucart, A., Debeir, O., Decaestecker, C., 2018. Artifact identification in digital pathology from weak and noisy supervision with deep residual networks. In: 2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech), pp. 1–6. https://doi.org/10.1109/CloudTech.2018.8713350.
Ganin, Y., Lempitsky, V., 2014. Unsupervised domain adaptation by backpropagation.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680
Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2018. HoVer-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. 10.48550/ARXIV.1812.06499.
Guha, I., Nadeem, S.A., You, C., Zhang, X., Levy, S.M., Wang, G., Torner, J.C., Saha, P.K., 2020. Deep learning based high-resolution reconstruction of trabecular bone microstructures from low-resolution CT scans using GAN-CIRCLE. Proc. SPIE Int. Soc. Opt. Eng. 11317.
Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. SMC-3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314.
Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2424–2433. https://doi.org/10.1109/CVPR.2016.266.
Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A., 2019. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin. Cancer Inform. 3, 1–7.
Jiang, Y.Q., Xiong, J.H., Li, H.Y., Yang, X.H., Yu, W.T., Gao, M., Zhao, X., Ma, Y.P., Zhang, W., Guan, Y.F., Gu, H., Sun, J.F., 2019. Recognizing basal cell carcinoma on smartphone-captured digital histopathology images with deep neural network. Br. J. Dermatol.
Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. Lect. Notes Comput. Sci. 694–711. https://doi.org/10.1007/978-3-319-46475-6_43.
Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining: formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data 6. https://doi.org/10.1145/2382577.2382579.
Kimeswenger, S., Tschandl, P., Noack, P., Hofmarcher, M., Rumetshofer, E., Kindermann, H., Silye, R., Hochreiter, S., Kaltenbrunner, M., Guenova, E., Klambauer, G., Hoetzenecker, W., 2020. Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns. Mod. Pathol.
Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250.
Mackiewicz-Wysocka, M., Bowszyc-Dmochowska, M., Strzelecka-Węklar, D., Dańczak-Pazdrowska, A., Adamski, Z., 2013. Basal cell carcinoma - diagnosis. Contemp. Oncol. (Pozn) 17, 337–342.
Marka, A., Carter, J.B., Toto, E., Hassanpour, S., 2019. Automated detection of nonmelanoma skin cancer using digital images: a systematic review. BMC Med. Imaging 19, 21.
Menéndez, M.L., Pardo, J.A., Pardo, L., Pardo, M.C., 1997. The Jensen-Shannon divergence. J. Franklin Inst. 334, 307–318. https://doi.org/10.1016/S0016-0032(96)00063-4.
Onajin, O., Wetter, D.A., Roenigk, R.K., Gibson, L.E., Weaver, A.L., Comfere, N.I., 2015. Frozen section diagnosis for non-melanoma skin cancers: correlation with permanent section diagnosis. J. Cutan. Pathol. 42, 459–464. https://doi.org/10.1111/cup.12498.
Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Comput. Graph. Appl. 21, 34–41. https://doi.org/10.1109/38.946629.
Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.
Samarasinghe, V., Madan, V., 2012. Nonmelanoma skin cancer. J. Cutan. Aesthet. Surg. 5, 3–10.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2016. Grad-CAM: visual explanations from deep networks via gradient-based localization.
Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2019. Staingan: stain style transfer for digital histological images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019. https://doi.org/10.1109/isbi.2019.8759152.
Sinn, M., Rawat, A., 2018. Non-parametric estimation of Jensen-Shannon Divergence in generative adversarial network training. In: Storkey, A., Perez-Cruz, F. (Eds.), Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. PMLR, pp. 642–651.
Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C., 2017. VEEGAN: reducing mode collapse in GANs using implicit variational learning.
Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2019. A closer look at domain shift for deep learning in histopathology.
Tampu, I.E., Eklund, A., Haj-Hosseini, N., 2022. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. Sci. Data 9, 580. https://doi.org/10.1038/s41597-022-01618-6.
Tellez, D., Litjens, G., B?ndi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.
Venables, Z.C., Autier, P., Nijsten, T., Wong, K.F., Langan, S.M., Rous, B., Broggio, J., Harwood, C., Henson, K., Proby, C.M., Rashbass, J., Leigh, I.M., 2019. Nationwide incidence of metastatic cutaneous squamous cell carcinoma in England. JAMA Dermatol. 155, 298–306. https://doi.org/10.1001/jamadermatol.2018.4219.
Wang, H.-.H., Wang, Y.-.H., Liang, C.-.W., Li, Y.-.C., 2019. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. JAMA Dermatol. 155, 1277–1283. https://doi.org/10.1001/jamadermatol.2019.2335.
Wei, Jerry, Suriawinata, A., Vaickus, L., Ren, B., Liu, X., Wei, Jason, Hassanpour, S., 2019. Generative image translation for data augmentation in colorectal histopathology images.
Wright, A.I., Dunn, C.M., Hale, M., Hutchins, G.G.A., Treanor, D.E., 2021. The effect of quality control on accuracy of digital pathology image analysis. IEEE J. Biomed. Health Inform. 25, 307–314. https://doi.org/10.1109/JBHI.2020.3046094.
Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.634.
Yanofsky, V.R., Mercer, S.E., Phelps, R.G., 2011. Histopathological variants of cutaneous squamous cell carcinoma: a review. J. Skin Cancer 2011, 210813.
You, C., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z., Cong, W., Vannier, M.W., Saha, P.K., Hoffman, E.A., Wang, G., 2020a. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). IEEE Trans. Med. Imaging 39, 188–203.
You, Chenyu, Yang, J., Chapiro, J., Duncan, J.S., 2020. Unsupervised wasserstein distance guided domain adaptation for 3D multi-domain liver segmentation. 10.48550/ARXIV.2009.02831.
Zhang, X., Karaman, S., Chang, S.-.F., 2019. Detecting and simulating artifacts in GAN fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). https://doi.org/10.1109/wifs47025.2019.9035107.
Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2015. Loss Functions for neural networks for image processing.
Zhu, J.-.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2017.244.