scientific reports

OPEN

Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images

Wingates Voon[1], Yan Chai Hum[1]✉, Yee Kai Tee[1], Wun-She Yap[2], Humaira Nisar[3], Hamam Mokayed[4], Neha Gupta[5] & Khin Wee Lai[6]

Debates persist regarding the impact of Stain Normalization (SN) on recent breast cancer histopathological studies. While some studies propose no influence on classification outcomes, others argue for improvement. This study aims to assess the efficacy of SN in breast cancer histopathological classification, specifically focusing on Invasive Ductal Carcinoma (IDC) grading using Convolutional Neural Networks (CNNs). The null hypothesis asserts that SN has no effect on the accuracy of CNN-based IDC grading, while the alternative hypothesis suggests the contrary. We evaluated six SN techniques, with five templates selected as target images for the conventional SN techniques. We also utilized seven ImageNet pre-trained CNNs for IDC grading. The performance of models trained with and without SN was compared to discern the influence of SN on classification outcomes. The analysis unveiled a *p*-value of 0.11, indicating no statistically significant difference in Balanced Accuracy Scores between models trained with StainGAN-normalized images, achieving a score of 0.9196 (the best-performing SN technique), and models trained with non-normalized images, which scored 0.9308. As a result, we did not reject the null hypothesis, indicating that we found no evidence to support a significant discrepancy in effectiveness between stain-normalized and non-normalized datasets for IDC grading tasks. This study demonstrates that SN has a limited impact on IDC grading, challenging the assumption of performance enhancement through SN.

Invasive ductal carcinoma (IDC) is widely recognized as the most common form of breast cancer, accounting for over 80% of breast cancer cases[1]. IDC grading is a crucial factor in determining the prognosis of IDC and plays a critical role in evaluating its clinical outcome. Henson et al.[2] found that the accuracy of IDC diagnosis improved when both the IDC grade and lymph node condition were considered. Similarly, the research conducted by Frkovic-Grazio and Bracko[3] demonstrated that IDC grading effectively predicts the behavior of the tumor, particularly for early-stage, small tumors. Schwartz et al.[4] also uncovered that when undergoing mastectomy, patients with high-grade IDC faced higher fatality rates and more frequent axillary lymph node involvement compared to those with lower-grade IDC. These findings highlight the significance of IDC grading in the prognostic evaluation of IDC.

The standard method of grading IDC is the Nottingham Grading Scheme (NGS), which is a semi-quantitative system based on three morphological features of IDC: mitotic count, nuclear pleomorphism, and degree of tubule formation[5]. These three criteria result in a total score that can be divided into Grades 1 to 3, which indicate the

[1]Department of Mechatronics and Biomedical Engineering, Faculty of Engineering and Science, Lee Kong Chian, Universiti Tunku Abdul Rahman, Kampar, Malaysia. [2]Department of Electrical and Electronic Engineering, Faculty of Engineering and Science, Lee Kong Chian, Universiti Tunku Abdul Rahman, Kampar, Malaysia. [3]Department of Electronic Engineering, Faculty of Engineering and Green Technology, Universiti Tunku Abdul Rahman, 31900 Kampar, Malaysia. [4]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Lulea, Sweden. [5]School of Electronics Engineering, Vellore Institute of Technology, Amaravati, AP, India. [6]Department of Biomedical Engineering, Universiti Malaya, 50603 Kuala Lumpur, Malaysia. ✉email: humyc@utar.edu.my

aggressiveness of the tumor. Lower-grade IDC is less aggressive, while higher-grade IDC is more aggressive[6]. Although manual IDC grading is still the standard, it can be time-consuming[7] and prone to high intra- and inter-observer variations, with agreement among pathologists reaching only 75.3% at best[8]. To address these limitations, automated IDC grading systems, a type of computer-aided diagnostic (CAD) technique, have been developed[9].

The development of automated IDC grading systems has significantly advanced from traditional handcrafted feature extraction methods[10–14] to the application of deep learning techniques[15–20]. This evolution extends beyond IDC grading, as deep learning also finds widespread utilization in various histopathological applications[21,22]. The process of generating digital IDC histopathological images involves several steps including the collection of IDC tissues, formalin fixation, paraffin section embedment, and staining with hematoxylin and eosin (H&E)[23,24]. The slides are then digitized using Whole Slide Imaging technology[25]. H&E staining, the standard protocol in histopathological studies, highlights cell nuclei in blue and different components such as cytoplasm and connective tissue with various shades of pink[26].

There is controversy surrounding the impact of Stain Normalization (SN) in recent breast cancer histopathological studies. Some studies have indicated that SN has no effect on classification results[26–30], while others have claimed that SN improves classification outcomes[31–35]. The purpose of SN is to address color inconsistencies in digital H&E-stained images caused by external factors such as the temperature of staining solutions, fixation characteristics, imaging device characteristics[36,37], and variations in light sources, detectors, or optics during slide digitization[38]. SN normalizes the color values of source images by matching the overall color distribution of target images[39]. However, the effectiveness of SN in improving classification results is uncertain due to the conflicting results in the literature.

In light of this, our aim is to investigate the effectiveness of SN in the breast cancer histopathological classification task using convolutional neural networks (CNNs), with a specific focus on classifying the Four Breast Cancer Grades (FBCG)[19] dataset into four IDC grades. We attempted to answer the question: "Is SN effective in the IDC grading task?" by conducting a statistical significance analysis using Student's *t*-test with the significance level, $\alpha = 0.05$. Below are our statements of null and alternative hypotheses:

1. Null hypothesis $\boldsymbol{H_0}$: A CNN trained with a stain-normalized dataset has no effect on the IDC grading accuracy.
2. Alternative hypothesis $\boldsymbol{H_1}$: A CNN trained with a stain-normalized dataset has an effect on the IDC grading accuracy.

In this paper, we selected six types of conventional and deep learning-based SN techniques to study their effectiveness with the IDC grading FBCG dataset using CNNs. The conventional methods, including Reinhard[40], Macenko[41], Structure-preserving Color Normalization (SPCN)[42] and Adaptive Color Deconvolution (ACD)[43] techniques, require a template as the stain target reference to stain-normalize the images. Hence, we selected five templates from the Patch Camelyon (PCam) dataset[44] (a dataset derived from Camelyon16[45]) for the conventional methods. For the deep learning-based SN methods, we utilized the Camelyon16[45] pre-trained StainGAN[46] and StainNet[47] to stain-normalize the images in the FBCG dataset. After normalizing the images, we implemented seven pre-trained CNNs: (1) EfficientNet-B0[48], (2) EfficientNet-V2-B0[49], (3) EfficientNet-V2-B0-21k[49], (4) ResNet-V1-50[50], (5) ResNet-V2-50[51], (6) MobileNet-V1[52], and (7) MobileNet-V2[53] as feature extractors in our IDC grading models to conduct the classification task. Our source code can be accessed publicly from: https://github.com/wingatesv/SN_IDC_Grading.

In this study, we have made the following contributions and reached the following conclusions:

1) We conducted a comprehensive evaluation of six conventional and deep learning-based SN techniques on the task of IDC grading using the FBCG dataset.
2) We conducted a systematic review of ten recent studies that investigated the efficacy of SN in breast cancer histopathological classification. The findings are presented in the section on related works.
3) Our results suggest that if SN is deemed necessary in the image pre-processing pipeline, StainGAN, StainNet, and ACD techniques are preferable to Reinhard, Macenko, and SPCN techniques.
4) Our statistical analysis revealed a *p*-value of 0.11 when comparing the mean balanced accuracy scores between models trained with the StainGAN-normalized FBCG dataset (the best performing SN technique), which achieved a score of 0.9196, and those trained with the non-normalized dataset, which scored 0.9308. This implies that we found no evidence of a significant difference in effectiveness between stain-normalized and non-normalized datasets for grading tasks.
5) Our findings challenge the assumption that stain normalization significantly improves histopathological classification tasks, as we found no evidence of a significant discrepancy in effectiveness between stain-normalized and non-normalized datasets for IDC grading tasks.

Our study provides insights into the effectiveness of SN techniques in breast cancer histopathological studies, with a particular focus on the IDC grading task. While there has been some debate over the impact of SN on classification outcomes, our research has shown that models trained with the non-normalized dataset can be just as effective as those trained with StainGAN-normalized images. Our findings provide a valuable contribution to the field and can help guide future research on SN techniques. We are optimistic that our study will encourage researchers to approach the topic with a critical lens and produce even more promising results in the future.

Related works

In this section, we examine the development of automated IDC grading systems and various SN methods. The SN techniques are divided into two categories: (1) conventional approaches and (2) deep learning-based approaches. Next, we present studies that investigated the effect of SN in various breast cancer histopathological image classification tasks.

Automated IDC grading systems

The development of automated IDC grading systems has progressed from manual feature extraction methods to deep learning-based approaches. For instance, Doyle et al.[11] proposed a method for extracting textural and architectural features by using spectral clustering to reduce the dimensionality of the extracted features, which were then used to classify the IDC grades with a support vector machine. Basavanhally et al.[13] employed a multifield-of-view (multi-FOV) classifier to identify the most salient image features from multiple FOV of varying sizes for the purpose of IDC grading. Dimitropoulos et al.[14] transformed images into vectors of locally aggregated descriptors (VLAD) representations based on the Grassmann manifold. They then calculated the VLAD encoding of each image on the manifold to determine the IDC grade. However, these methods are heavily reliant on features and are computationally intensive, with a lack of heuristics for feature extraction[18]. As a result, more recent studies have shifted towards deep learning methods, specifically Convolutional Neural Networks (CNNs)[15,17,18,54]. For example, Senousy et al.[18] developed an entropy-based elastic ensemble of CNNs (3E-Net) for IDC grading, and Yan et al.[55] created a nuclei-guided network (NGNet) with a nuclei-guided attention module for IDC grading as well. In terms of transfer learning, Zavareh et al.[20] used the VGG16 model as a feature extractor in the BCNet to grade IDC. Similarly, Voon et al.[56] evaluated the performance of seven pre-trained CNN models in the IDC grading task. In this study, we adopted the model implementation of Voon et al.[56] which utilized transfer learning. This approach was chosen due to the improved performance of CNNs when trained on a limited number of training images.

Stain normalization methods

Conventional stain normalization methods

Conventional approaches to Stain Normalization (SN) in histopathological images typically involve the analysis, transformation, and alignment of the color components of images[47]. The Reinhard method[40] normalizes the images by adjusting the statistical color distribution of the source image to match that of a template image while preserving the background color and color intensities. The Macenko technique[41] employs Single Value Decomposition (SVD) to form a plane that projects information, determining the corresponding angles and finally estimating the color matrix. The Khan method[39] identifies the stain color of the source image using the Stain Color Descriptor (SCD), then uses a Relevance Vector Machine (RVM) to determine the position of each stain and transfers the color from the template to the source image using a non-linear spline-based color normalization technique. The Structure-Preserving Color Normalization (SPCN)[42] decomposes the source images into sparse stain density maps, combining the stain of the template image to change only the color while preserving the structures. The Adaptive Color Deconvolution (ACD)[43] normalizes the stains by integrating optimization to approximate the parameters of stain separation and color normalization. This technique, based on Color Deconvolution (CD)[57], optimizes the estimation of stain parameters. However, these methods[39–43] depend on a reference image to approximate the stain parameters, presenting a challenge to encompass all staining patterns or represent all input images. As a result, the use of suboptimal reference images may lead to incorrect estimation of stain parameters and result in inaccurate outcomes[58,59].

Deep learning-based stain normalization methods

Recently, a significant shift has been observed towards the adoption of deep learning-based techniques for stain normalization (SN). This approach offers a departure from traditional methods that rely on template images[46,47,60,61]. Zanjani et al.[60] proposed the use of generative adversarial networks (GANs) to learn the relationship between image content structures and their respective color attributes, thereby facilitating color alignment without relying on statistical properties. Shaban et al.[46] extended this work by developing StainGAN, a CycleGAN-based technique[62] that enables the transfer of stain style from one domain to another without the need for paired data. Similarly, Kang et al.[47] introduced StainNet, a method that leverages the output of StainGAN to better understand the pixel-wise color mapping relationship within a given dataset. In our current study, we sought to investigate the effectiveness of SN in the context of IDC grading. To this end, we considered a diverse range of techniques, including Reinhard, Macenko, Structure-Preserving Color Normalization (SPCN), Adaptive Color Deconvolution (ACD), StainGAN, and StainNet.

Study of stain normalization in breast cancer histopathological images classification

This section presents an overview of the prior studies that have compared the performance of models trained with stain-normalized and non-normalized inputs in the context of breast cancer histopathological image classification. Despite the numerous studies in this field, there is still considerable controversy regarding the efficacy of SN on the performance of these models[26–35].

On one hand, several studies[26–30] have reported that SN has no significant impact on the performance of the models. For example, Gupta et al.[27] evaluated the classification performance of different texture descriptors and contemporary classifiers using Reinhard-normalized BreaKHis[63] dataset and found that SN did not lead to improvement in the results. Similarly, Tellez et al.[26] compared the performance of CNNs trained on Camelyon17[64] dataset using Macenko and Berjnodi[36] SN techniques, and revealed that SN did not enhance the performance, with the CNN trained on the non-normalized dataset even outperforming those trained on the stain-normalized datasets. These findings were supported by Kumar et al.[28], who found that a pre-trained VGG16 model trained on the non-normalized BreaKHis dataset outperformed the identical model trained on the Macenko-normalized dataset. Hameed et al.[29] also found that the performance of deep learning-based ensemble models declined when using stain-normalized datasets, while Hameed et al.[30] failed to find any performance improvement when the pre-trained Xception model was trained on the Colsanitas dataset[29] with Reinhard, Macenko, CD, and SPCN SN techniques.

On the other hand, several studies[31–35] have suggested that SN does indeed improve the performance of the models. For example, Nawaz et al.[31] fine-tuned the AlexNet model on the ICIAR2018 dataset[65] and found that the AlexNet trained on the Macenko-normalized dataset outperformed the model trained on the non-normalized dataset. Shahidi et al.[35] compared the performance of different CNNs on Macenko-normalized and non-normalized BreaKHis datasets and found that SN improved the model performance. Munien and Viriri[32] implemented seven pre-trained EfficientNets to classify the original, Reinhard-normalized, and Macenko-normalized ICIAR2018 datasets. The results showed that models trained with stain-normalized datasets outperformed models trained with the non-normalized dataset. Salvi et al.[33] attempted to classify the BACH challenge[65] dataset with Stain Color Adaptive Normalization (SCAN) technique[66]. The authors found that the normalized dataset obtained better results than the non-normalized dataset. Similarly, Alkassar et al.[34] utilized an ensemble of models to classify Khan-normalized and non-normalized BreaKHis datasets. The results showed that the models trained with the Khan-normalized dataset outperformed those trained with the non-normalized dataset. Therefore, we can conclude that these studies highlighted the benefits of SN in the classification task.

These inconsistent findings have created a knowledge gap in the application of SN in IDC grading, leading to confusion among researchers about the effectiveness of SN in future studies. In light of this, we set out to answer the question: "Is SN effective in the IDC grading task?" by investigating the effectiveness of six conventional and deep learning-based SN techniques on the IDC grading task using the FBCG dataset and CNNs.

Methodology

Overview

In this section, we provide an outline of the six SN techniques used in the IDC grading task. We also elucidate the implementation details, which include aspects such as the FBCG dataset, image pre-processing procedures, CNN model implementations, and the evaluation metric. All experimentations were conducted using Python and TensorFlow Keras on the Google Collaboratory platform. The technical specifications for these experiments included a 2.30 GHz Intel® Xeon® CPU, up to 32 GB RAM, and an NVIDIA P100 or T4 GPU. We ensure that all procedures adhered to relevant guidelines and regulations. Figure 1 illustrates the general methodology of the study.

Stain normalization

SN aims to normalize the color values of the source images by aligning the overall color distribution with that of target images. Our study explored six types of SN techniques, specifically Reinhard[40], Macenko[41], SPCN[42], ACD[43], StainGAN[46] and StainNet[47] (Note that the employed StainGAN and StainNet were pre-trained on the Camelyon16 dataset[45]).

Template selection

The selection of an appropriate template is crucial for conventional SN techniques, which rely on a single template to perform color conversion between source and target images. If the template is not chosen wisely, the performance of SN techniques may be compromised[47]. Therefore, we selected five templates where $T \in \{T1, T2, T3, T4, T5\}$ (see Fig. 2) from the PatchCamelyon (PCam) dataset[44], our target dataset, to investigate the impact of each template on the SN techniques. It is imperative to note that the selection of these templates was not selected based on subjective decisions. Instead, they were chosen based on a methodical process that involved generating an average image from the target dataset and using similarity functions to compare this average image with image samples within the target dataset. This approach helped us identify a template that most accurately reflects the overall color staining distribution of the dataset.

Average image generation. Before selecting any template, we generated an image I_{avg} that represents the average pixel values of the target dataset. In this case, we selected the PCam train set as the target dataset D_t to ensure a fairer comparison with the StainGAN and StainNet SN techniques. PCam is a public histological dataset which comprises patch-wise images with dimensions of 96 by 96 pixels. These images are extracted from histological scans of lymph node sections from the Camelyon16 Challenge, which focuses on breast cancer metastasis. To generate I_{avg}, all 262,144 images from the PCam train set were converted into floating-point arrays, followed by summing up the arrays to yield the average pixel values.

Templates 1 and 2. Template 1 ($T1$) and Template 2 ($T2$) were selected using cosine similarity SIM_C. This method computes the dot product of two vectors and divides it by the product of their magnitude to determine their similarity. Specifically, we computed the SIM_C between I_{avg} and image $X \in D_t$ to locate X that most resembles I_{avg}, resulting in $T1$. Likewise, selecting $T2$ adopted a similar approach. However, the most dominant color, C_{dom} of I_{avg} and image $X \in D_t$ were obtained, followed by forming image $I_{avg,dom}$ and $I_{Dt,dom}$ based on each dominant color, respectively. Subsequently, we computed the SIM_C between $I_{avg,dom}$ and $I_{Dt,dom}$, resulting in $T2$. Equation (1) formally describes the SIM_C:

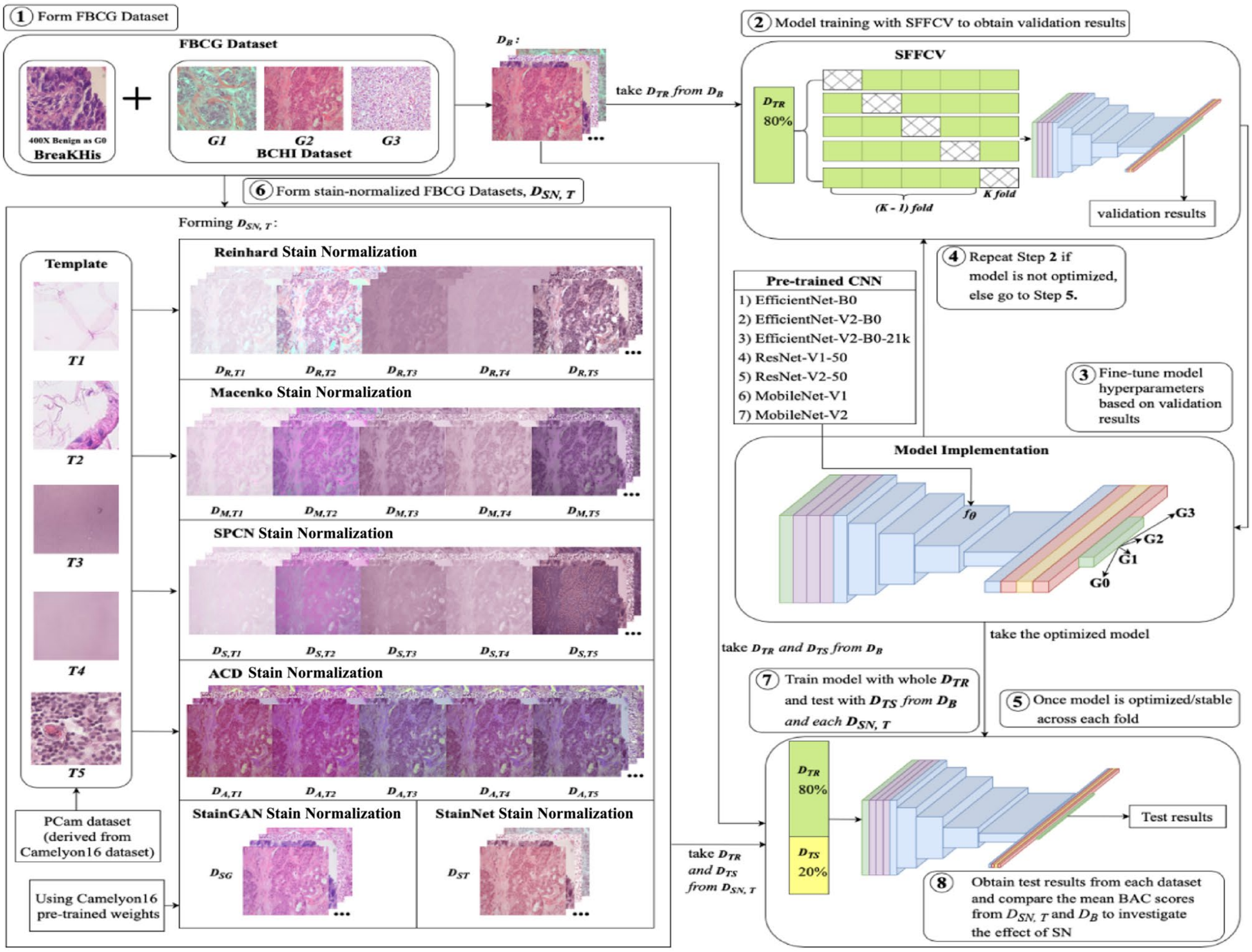
1 Form FBCG Dataset
FBCG Dataset
400X Benign as G0
BreaKHis
G1
G2
G3
BCHI Dataset
D_B:
take D_{TR} from D_B
2 Model training with SFFCV to obtain validation results
SFFCV
D_{TR}
80%
(K - 1) fold
K fold
validation results
4 Repeat Step 2 if model is not optimized, else go to Step 5.
Pre-trained CNN
1) EfficientNet-B0
2) EfficientNet-V2-B0
3) EfficientNet-V2-B0-21k
4) ResNet-V1-50
5) ResNet-V2-50
6) MobileNet-V1
7) MobileNet-V2
3 Fine-tune model hyperparameters based on validation results
Model Implementation
$f\theta$
G3
G2
G1
G0
take the optimized model
take D_{TR} and D_{TS} from D_B
7 Train model with whole D_{TR} and test with D_{TS} from D_B and each $D_{SN,T}$
5 Once model is optimized/stable across each fold
D_{TR}
80%
D_{TS}
20%
Test results
8 Obtain test results from each dataset and compare the mean BAC scores from $D_{SN,T}$ and D_B to investigate the effect of SN
take D_{TR} and D_{TS} from $D_{SN,T}$
6 Form stain-normalized FBCG Datasets, $D_{SN,T}$
Forming $D_{SN,T}$:
Template
T1
T2
T3
T4
T5
PCam dataset (derived from Camelyon16 dataset)
Using Camelyon16 pre-trained weights
Reinhard Stain Normalization
$D_{R,T1}$
$D_{R,T2}$
$D_{R,T3}$
$D_{R,T4}$
$D_{R,T5}$
Macenko Stain Normalization
$D_{M,T1}$
$D_{M,T2}$
$D_{M,T3}$
$D_{M,T4}$
$D_{M,T5}$
SPCN Stain Normalization
$D_{S,T1}$
$D_{S,T2}$
$D_{S,T3}$
$D_{S,T4}$
$D_{S,T5}$
ACD Stain Normalization
$D_{A,T1}$
$D_{A,T2}$
$D_{A,T3}$
$D_{A,T4}$
$D_{A,T5}$
StainGAN Stain Normalization
D_{SG}
StainNet Stain Normalization
D_{ST}

Figure 1. The overall methodology of the study. (1) The FBCG dataset is assembled by combining images from the 400X Benign class of the BreaKHis dataset and images from the BCHI dataset. (2) To evaluate model stability, the implemented model is trained with D_{TR} from D_B using the Stratified Five-fold Cross-validation (SFFCV). (3) The hyperparameters of the model are optimized until the model is stable across each fold. (4) The SFFCV process is repeated until the model is optimized. (5) Once satisfactory model performance is achieved, (6) the FBCG datasets undergo stain normalization using various techniques to form $D_{SN,\,T}$. (7) Lastly, each $D_{SN,\,T}$ and D_B is fed forward into the model to retrain, followed by (8) obtaining the final test results.

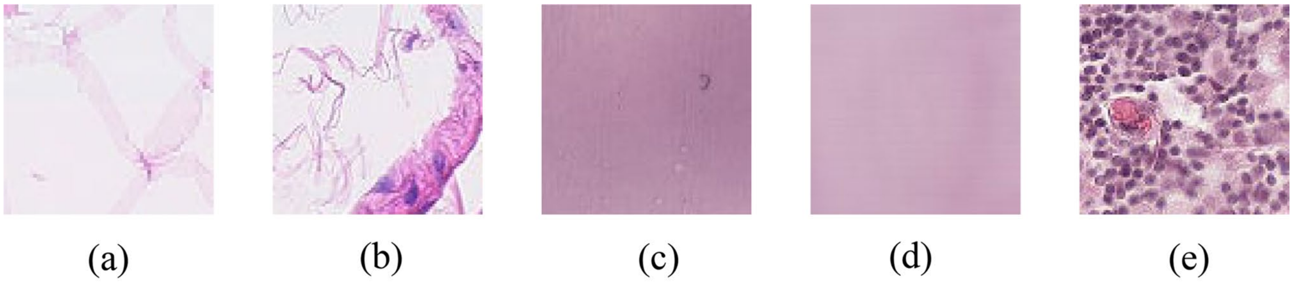

Figure 2. Five templates selected from PCam train set: (**a**) $T1$, (**b**) $T2$, (**c**) $T3$, (**d**) $T4$ and (**e**) $T5$.

$$SIM_C(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

where A and B denote vectors with $n\text{-}th$ number of pixels flatten from I_{avg} and image $X \in D_t$ or $I_{avg,dom}$ and $I_{Dt,dom}$. Equation (2) formally describes the C_{dom}:

```
Input: $I_s, T \in \{T1, T2, T3, T4, T5\}$
Output: $I_{out}$
Initiate $i = 0$, number of channel, $c = 3$
Read $I_s$ and $T$
Convert $I_s$ and $T$ from $RGB$ to $l\alpha\beta$ colour space
while $i < c$ do
    $I_{out} \leftarrow$ transform $I_s$ with Equations (5), (6) and (7)
    $i \leftarrow i + 1$
end while
convert $l\alpha\beta$ back to $RGB$ color space
```

Algorithm 1. Reinhard Technique

$$C_{dom} = \arg\max_{c \in C(P)} N(c) \tag{2}$$

where *P* denotes the set of all pixels in an image, *C(p)* denotes the function that returns the color of pixel *p*, and *N(c)* denotes the function that returns the number of pixels of color *c* in the image.

Templates 3, 4 and 5. For Templates 3, 4, 5, we used different selection methods. Template 3 ($T3$) was selected using the Mean Square Error *MSE*, while Template 4 ($T4$) was chosen based on the Structural Similarity Index *SSIM*. Similar to $T1$ and $T2$, we computed the *MSE* and *SSIM* between I_{avg} and image $X \in D_t$ to find the most similar *X*, resulting in $T3$ and $T4$. For Template 5 ($T5$), we identified the most dominant color in I_{avg} and image $X \in D_t$, We then formed images $I_{avg,dom}$ and $I_{Dt,dom}$, based on each dominant color. Then, we computed the *MSE* or *SSIM* between $I_{avg,dom}$ and $I_{Dt,dom}$, resulting in $T5$ (note that the results of *MSE* and *SSIM* are identical). Equations (3) and (4) describe *MSE* and *SSIM* respectively as followed:

$$MSE(I_A, I_B) = \frac{1}{n}\sum_{i=1}^{n}(I_{A,i} - I_{B,i})^2 \tag{3}$$

$$SSIM(I_A, I_B) = \frac{(2\mu_{IA}\mu_{IB} + C_1)(2\sigma_{IAIB} + C_2)}{(\mu_{IA}^2 + \mu_{IB}^2 + C_1)(\sigma_{IA}^2 + \sigma_{IB}^2 + C_2)} \tag{4}$$

where I_A and I_B denote input and output image matrices with *n-th* number of pixels respectively, μ_{IA} and μ_{IB} denote the luminance of I_A and I_B respectively, σ_{IA} and σ_{IB} denote the contrast of I_A and I_B respectively, C_1 and C_2 denote constants to ensure stability where C_1 and $C_2 > 0$.

Reinhard stain normalization technique

The Reinhard SN technique normalizes the source image I_s by aligning the mean μ and standard deviation σ with a template *T*. Algorithm *1* outlines the workflow of the Reinhard algorithm. The Reinhard method transforms the *RGB* images to $l\alpha\beta$ color space where *l* represents the achromatic channel, α denotes the chromatic blue-yellow channel and β signifies the chromatic green–red channel. Subsequently, the following Eqs. (5), (6) and (7) are applied to perform the Reinhard transformation, then convert the output image I_{out} back to *RGB* color space[40,68].

$$l_2 = \mu(l_1) + (l - \mu(l)) \odot (\sigma(l_1) \oslash \sigma(l)) \tag{5}$$

$$\alpha_2 = \mu(\alpha_1) + (\alpha - \mu(\alpha)) \odot (\sigma(\alpha_1) \oslash (\alpha)) \tag{6}$$

$$\beta_2 = \mu(\alpha_1) + (\beta - \mu(\beta)) \odot (\sigma(\beta_1) \oslash \sigma(\beta)) \tag{7}$$

where l, l_1 and l_2 depict the I_S, T and I_{out} in the *l* space respectively; α, α_1 and α_2 depict the I_S, T and I_{out} in the α space respectively; β, β_1 and β_2 depict the I_S, T and I_{out} in the β space respectively; $\odot$ denotes element-wise multiplication and $\oslash$ denotes element-wise division.

Macenko stain normalization technique

The Macenko technique separates stains by identifying the fringe of pixel distribution in the Optical Density space (*OD*). Algorithm *2* provides a detailed description of the Macenko algorithm. Similar to Reinhard, Macenko converts the *RBG* image to $l\alpha\beta$ color space, followed by transforming the colors into *OD* values with Eq. (8):

$$OD = -log_{10}(I_s) \tag{8}$$

The color transformation to *OD* values provides a space where a linear stain fusion yields a linear fusion of *OD* values. Subsequently, the transparent pixels are removed if the *OD* value is below a specific threshold. The *OD* value is split into two matrices, given by Eqs. (9) and (10).

```
Input: $I_s, T \in \{T1, T2, T3, T4, T5\}$
Output: $I_{out}$
Initiate tolerance for the pseudo-minimum $\alpha^{th}$ and pseudo-maximum $(100 - \alpha)^{th}$
percentile, where $\alpha = 1$, $OD$ threshold value for transparent pixels $\beta = 0.15$,
transmitted light intensity $I_o = 240$
Read $I_s$ and $T$
Convert $I_s$ and $T$ from $RGB$ to $OD$ colour space with Equation (8)
if $OD < \beta$ then
  remove transparent pixels
else
    Compute $SVD$ on the $OD$
    Devise plane from the $SVD$ directions
    Project data onto the plane, normalizing to unit length
    Compute each angle point corresponding to the first $SVD$ direction
    Locate robust extremes and transform extreme values back to $OD$ space
    Determine normalization stain concentration
    $I_{out} \leftarrow r$ecreate the normalized image using reference mixing matrix
end if
```

Algorithm 2. Macenko Technique

$$OD = V * S \tag{9}$$

$$S = V' * OD \tag{10}$$

where *S* represents each stain saturation and *V* denotes stain vector matrix. Equations (8) and (9) locate the stain vector of each image based on the color (if *OD* = 0, then the corresponding pixel = white; the stain is absent). Next, we compute the singular value decomposition (SVD) on the *OD* value, followed by locating the stain vector terminal points using the Geodesic path[37]. We can then assess the plane, which is created by vectors. The procedure is conducted by creating a plane with two vectors corresponding to the two most significant *SVD* values. Afterwards, we project all *OD* values into the plane, normalizing to unit length and curving the projected line. With these, we can compute each angle to the first *SVD* direction, thus, mapping the direction in the plane. As a result, the pixel intensity histogram can be computed, followed by determining the concentration of each stain with the H&E matrix in relation to the *OD* values. Finally, we can yield I_{out} by using the H&E matrix with the normalized stain concentration[41,68].

Structure-preserving color normalization

Structure-Preserving Color Normalization (SPCN)[42] operates by decomposing I_s into sparse stain density maps while integrating the stain from *T*. Algorithm *3* illustrates the implementation of SPCN. Given $I \in \mathbb{R}^{m \times n}$ is the *RGB* image matrix, where *m* denotes the number of *RGB* channels and *n* denotes the number of pixels. Let $w \in \mathbb{R}^{m \times r}$ be the stain matrix with columns representing the chromatic variance of each stain, where *r* represents the stain number. Let $H \in \mathbb{R}^{r \times n}$ represents the stain density maps where the rows denote the stain concentration. Thus, *I* is described as:

$$I = I_o e^{-WH} \tag{11}$$

Let *V* be the OD maps then,

$$V = log(\frac{I_o}{I}) \tag{12}$$

By utilizing Eq. (11), we can form:

$$V = WH \tag{13}$$

where *V* = observation matrix, *H* = stain density map matrix, and *W* = stain color appearance matrix. Next, we implement the sparse non-negative matrix factorization (SNMF) for stain separation. Based on the Beer-Lambert law, the *RGB* image is converted into the *OD* maps with Eq. (14). Then, the sparseness constraint is added in Eq. (11). SNMF separates stain with l_1= sparseness and H_j = stain mixing coefficient where, *j* = index of stains that is $j = 1, 2, \ldots\ldots r,$.

$$\varphi(p) = -\log(V(p)) \tag{14}$$

where φ denotes as the OD space, *p* = pixel intensity where, $p \in pixelP$.

```
Input: $I_s, T \in \{T1, T2, T3, T4, T5\}$
Output: $I_{out}$
Read $I_s$ and $T$
Apply $BLT$ with Equation (14)
Sparse stain separation using $SNMF$ with Equations (15) and (16)
Stain normalization
Apply inverse $BLT$
$I_{out} \leftarrow$ normalized $I_s$
```

Algorithm 3. Structure-Preserving Color Normalization

$$min\frac{1}{2}\|V - WH\|_F^2 + \lambda\sum\nolimits_{j=1}^{r}\|H(j,:)\|_1, W, H \geq 0 \tag{15}$$

$$\|W(:,j)\|_2^2 = 1 \tag{16}$$

where λ = the sparsity and regularization parameter. Additional constraints on W and H will decrease the solution space of W/α and αH, where α is a positive value. Equation (12) represents a non-convex optimization problem, which can be addressed by alternating optimizing one parameter of H and W while holding the other constant. Elements are randomly selected from the optical density V to initialize the color appearance matrix.

Subsequently, we transfer the color μ of T to I_s while approximating the color appearance matrix for stain normalization. Utilizing the SNMF, we factorize the stain density maps V_s into W_sH_s and V_t into W_tH_t. Afterwards, the stain density maps of source H_s are merged with the template W_t color appearance matrix instead of the source color appearance matrix W_s to produce the normalized image. As a result, stain density map H maintains the structure while the color appearance matrix W maintains changes in the color appearance. Lastly, the inverse Beer-Lambert transformation (BLT) is applied to the normalized stains to obtain I_{out}[42,68].

Adaptive color deconvolution technique

Adaptive Color Deconvolution (ACD)[43] normalizes stains by integrating optimization to approximate the stain separation parameters and color normalization. ACD is based on color deconvolution (CD)[57]. Let $x_i \in \mathbb{R}^{3\times1}$ denote the RGB values of each i-th pixel in I_s. CD is described with Eqs. (17) and (18):

$$o_i = -\ln(\frac{x_i}{I_{max}}) \tag{17}$$

$$s_i = D \cdot o_i \tag{18}$$

Where $o_i \in \mathbb{R}^{3\times1}$ represents the OD of RGB channels, I_{max}= background intensity, and $D \in \mathbb{R}^{3\times3}$ = CD matrix. The separated densities of stains are denoted $s_i = (h_i, e_i, d_i)^T$, where h_i = hematoxylin stain, e_i = eosin stain, and d_i = separation residual. CD matrix D is decided by a Stain Color Appearance (SCA) matrix M, where $D = M^{-1}$. Therefore, ACD is derived by applying a stain-weight matrix $W = diag(w_h, w_e, 1)$ to directly optimize the stain separation parameters and color normalization. We modify Eq. (18) to form Eq. (19):

$$s_i = W \cdot D \cdot o_i \tag{19}$$

The SCA matrix $M = (m_h, m_e, m_d)$, where $m_j \in \mathbb{R}^{3\times1}(j = h, e, d)$ is a unit vector representing the contributions of the j-th stain to the RGB channels intensities. M is determined by φ, representing as M (φ) and CD matrix D as D (φ), where φ is a collection of six-degree variables $\varphi = \{\alpha_h, \beta_h, \alpha_e, \beta_e, \alpha_d, \beta_d\}$. Thus, we perform optimization by minimizing the objective function $\mathcal{L}_{ACD}$[43] of variables φ and W:

$$(\widehat{\varphi}, \widehat{W}) = \text{argmin}\mathcal{L}_{ACD}(\varphi, W) \tag{20}$$

We employed the gradient descent to solve $\mathcal{L}_{ACD}(\varphi, W)$ which is continuous and differentiable for variables φ and W. By resolving $\mathcal{L}_{ACD}$, $\widehat{\varphi}$ and $\widehat{W}$ can be obtained, followed by determining the adaptive matrices M ($\widehat{\varphi}$) and D ($\widehat{\varphi}$) for the I_s. After the optimization, we obtain the adaptive variables for the stain separation $\widehat{D}$ and stain intensity normalization $\widehat{W}$. Subsequently, we separate the I_s stain components with $\widehat{D}$, followed by weighting with $\widehat{W}$. Lastly, we recombine the weighted stain components with the SCA matrix of the template T $\overline{M}$ to obtain I_{out}. The following Eqs. (17), (21) and (22) summarize ACD techniques for the i-th pixel x_i:

$$o_i = -\ln(\frac{x_i}{I_{max}}) \tag{17}$$

$$\overline{o_i} = \overline{M} \cdot \widehat{W}\widehat{D} \cdot o_i \tag{21}$$

```
Input: $I_s$, $T \in \{T1, T2, T3, T4, T5\}$, $I_{max}$, $D$, $\varphi$, $\overline{M}$
Output: $I_{out}$
Read $I_s$ to obtain $x_i$
Read $T$ to obtain $\overline{M}$
Convert $RGB$ to $OD$ space with Equation (17)
$D\ (\hat{\varphi}), \widehat{W} \leftarrow$ optimise Equation (19) by minimizing $\mathcal{L}_{ACD}(\varphi, W)$ with gradient descent
$I_{out} \leftarrow$ stain separation, weighting $\widehat{W}$ and recombination with $\overline{M}$ with Equations (21) and (22)
Convert $OD$ back to $RGB$ color space
```

Algorithm 4. Adaptive Color Deconvolution

$$I_{out} = \overline{x_i} = \exp(-\overline{o_i}) \cdot I_{max} \tag{22}$$

StainGAN

StainGAN[46] is inspired by CycleGAN[62] that transfers stains between two domains without requiring paired data from both domains. StainGAN is composed of two pairs, each consisting of a generator and a discriminator. The first pair (G_A and D_A) aims to map images from Domain B to Domain A $G_A : X_B \rightarrow X_A$. The Generator G_A aims to generate images that match Domain A. The discriminator D_A tries to verify if images originate from Domain A or the fake generated ones. The other pair (G_B and D_B) undergoes the same process in the reverse direction, $G_B : X_A \rightarrow X_B$ as:

$$\widehat{X_A} = G_A(X_B;\theta_A), \widehat{X_B} = G_B(X_A;\theta_B), s.t.d\left(X_B, \widehat{X_B}\right) \leq \epsilon \tag{23}$$

$$\widehat{X_B} = G_B(X_A;\theta_B), \widehat{X_A} = G_A(X_B;\theta_A), s.t.d\left(X_A, \widehat{X_A}\right) \leq \epsilon \tag{24}$$

where $d\ (\cdot,\cdot)$ = distance metric between the input image and the reconstructed image (cycle-consistency constraint), and both θ_A and θ_H are the model parameters. StainGAN is trained to minimize adversarial and cycle-consistency loss (see Algorithm *5* for StainGAN training details). The cycle-consistency loss ensures that the output from G_A can be reconstructed back to the input for G_B, and similarly, the output from G_B can be reconstructed back to the input for G_A. The adversarial loss assures that the stain of the reconstructed images is coherent with the actual stain distribution.

Where the cycle-consistency loss for the $B \rightarrow A \rightarrow B$ cycle, $\mathcal{L}_{cycle}^{(B\rightarrow A\rightarrow B)}$ is described as follow:

$$\mathcal{L}_{cycle}^{(B\rightarrow A\rightarrow B)} = \frac{1}{m}\sum_{i=1}^{m}(b^{(i)} - D_{A\rightarrow B}\left(G_{B\rightarrow A}(b^{(i)})\right))^2 \tag{25}$$

StainNet

StainNet[47] normalizes the source dataset by learning the color mapping relationship from the target dataset and adjusting its color value pixel by pixel. StainNet is a CNN comprising three convolutional layers with 32 kernels. StainNet necessitates the pairing of source and target images to facilitate the learning of color space conversion from the source to the target. Therefore, StainNet relies on the output of StainGAN to obtain the paired images. Specifically, we treat StainGAN as the teacher model while StainNet as the student model. The output images from StainGAN are treated as truth labels for the StainNet to train. Thus, the primary objective of the StainNet is to minimize the *L1* loss with *SGD* optimizer corresponding to the normalized images generated by StainGAN (see Algorithm *6* for StainNet training details). The mapping association of StainGAN is contingent on the image content. Therefore, by training on images normalized by StainGAN, StainNet can convert the content-based mapping association of StainGAN into a pixel value-based mapping.

Implementation details

This section outlines the implementation details of training CNN models on various stain-normalized datasets. The objective is to evaluate the performance of these models when trained on diverse stain-normalized datasets.

Dataset description

FBCG dataset. We adopted the dataset strategy proposed by Abdelli et al.[19], known as the Four Breast Cancer Grades (FBCG) dataset to address the limitations of the existing small IDC grading datasets. The FBCG dataset entails 888 RGB H&E stained 400X-magnification IDC histopathological images with four classes: Grade 0 (G0),

```
Inputs: Domain $A$, Domain $B$
for $epoch \in epochs$ do
    Draw a minibatch of samples $\{a^{(1)},...,a^{(m)}\}$ from Domain $A$
    Draw a minibatch of samples $\{b^{(1)},...,b^{(m)}\}$ from Domain $B$
    Compute discriminator loss on inputs from Domain $A$:
        $\mathcal{L}_A^{(D)} = \frac{1}{m}\sum_{i=1}^{m}(D_A(a^{(i)})-1)^2 + \frac{1}{n}\sum_{j=1}^{n}(D_B(b^{(j)})-1)^2$
    Compute discriminator loss on inputs from Domain $B$:
        $\mathcal{L}_B^{(D)} = \frac{1}{m}\sum_{i=1}^{m}\left(D_B\left(G_{A\to B}(a^{(i)})\right)\right)^2 + \frac{1}{n}\sum_{j=1}^{n}\left(D_A\left(G_{B\to A}(b^{(j)})\right)\right)^2$
    Update $D_A$ and $D_B$
    Compute the $B\to A$ generator loss:
        $\mathcal{L}^{(G_{B\to A})} = \frac{1}{m}\sum_{i=1}^{m}\left(D_A\left(G_{B\to A}(b^{(i)})\right)-1\right)^2 + \mathcal{L}_{cycle}^{(B\to A\to B)}$
    Compute the A $\to B$ generator loss:
        $\mathcal{L}^{(G_{A\to B})} = \frac{1}{n}\sum_{j=1}^{n}\left(D_B\left(G_{A\to B}(a^{(j)})\right)-1\right)^2 + \mathcal{L}_{cycle}^{(A\to B\to A)}$
    Update $G_A$ and $G_B$
end for
```

Algorithm 5. StainGAN Training Loop

```
Input: normalized images from StainGAN, $D_{SG}$
for $epoch \in epochs$ do
    for $X\ batch \in D_{SG}$ do
        $\hat{y}$ = StainNet $(X, \theta)$
        Compute $loss = L1_{loss}(\hat{y}, y)$
        Compute gradient, $\nabla_{loss}$ of the $\theta$ with respect to the $loss$
        Update $\theta \leftarrow SGD(\nabla_{loss}, \theta)$
    end for
end for
```

Algorithm 6. StainNet Training Loop

		Grade 0	Grade 1	Grade 2	Grade 3	Total
FBCG Dataset	Train set	470	86	82	73	711
	Test set	118	21	20	18	177
	Total	588	107	102	91	888

Table 1. The class distribution and proposed train-test split of the FBCG dataset.

Grade 1 (G1), Grade 2 (G2), and Grade 3 (G3). The images in the G0 class (588 in total) are sourced from the Benign class of the BreaKHis dataset[63], captured at a 400X magnification. The images in the other classes (300 in total) are sourced from the BCHI dataset[69]. Table 1 summarizes the composition of the FBCG dataset.

BCHI dataset. The Breast Carcinoma Histological Images (BCHI) dataset[69] includes 300 H&E-stained breast histopathology images (1280 × 960 pixels) from the pathology department at "Agios Pavlos" Hospital in Thessaloniki, Greece. The images, which depict carcinoma specimens, are categorized into three grades: Grade 1 (with 107 images), Grade 2 (with 102 images), and Grade 3 (with 91 images). These images are sourced from 21 IDC

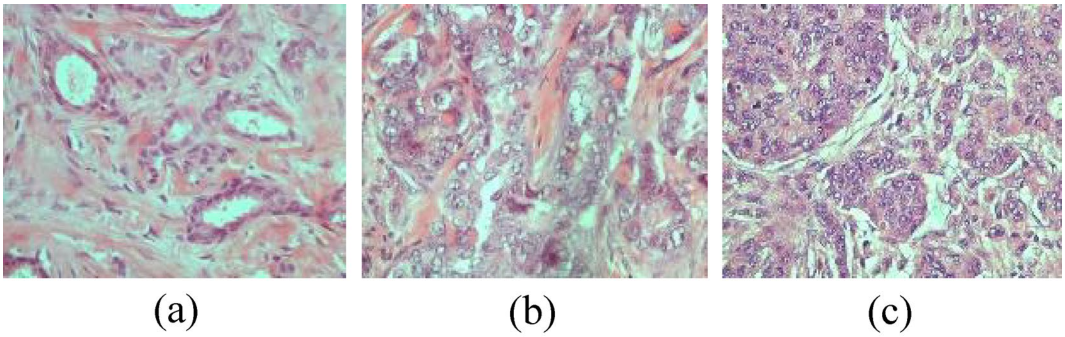

Figure 3. Samples images with 400X magnification from the BCHI dataset: (**a**) Grade 1, (**b**) Grade 2, (**c**) Grade 3.

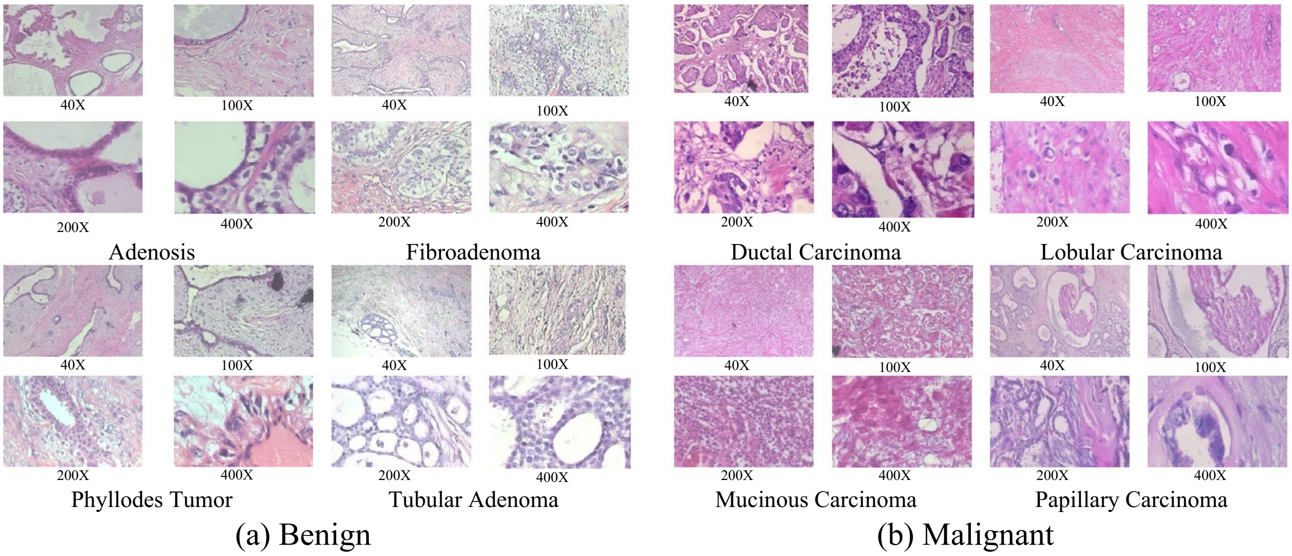

Figure 4. Samples from the BreaKHis dataset distributed into two major classes: (**a**) Benign and (**b**) Malignant with four magnification factors.

patients. The images were captured using a Nikon camera and a 40X magnification objective lens on a compound microscope (see Fig. 3).

BreaKHis dataset. The BreaKHis dataset[63] comprises 7909 histopathological images of breast cancer, sourced from 82 patients. Initially, the H&E-stained slide was captured at four magnification factors (40X, 100X, 200X, and 400X), using four objective lenses (4X, 10X, 20X, and 40X). These images were then converted into digital RGB format dimensions of 700 by 460 pixels. The BreaKHis is primarily divided into two categories: (1) Benign (2480 images) and (2) Malignant (5429 images). Each of the category can be further subdivided into four subclasses. For the Benign class, these are: (1) Adenosis, (2) Fibroadenoma, (3) Phyllodes Tumor, and (4) Tubular Adenoma. For the Malignant class, the subclasses are: (1) Ductal Carcinoma, (2) Lobular Carcinoma, (3) Mucinous Carcinoma, and (4) Papillary Carcinoma (see Fig. 4). Table 2 provides a detailed distribution of the images by major classes and magnifications within the BreaKHis dataset.

Experiment setup

In this study, we assessed the base dataset (original FBCG dataset), represented as D_B, comprising $2D$ pixel elements with three RGB channels and their corresponding ground truth labels. We employed six selected SN techniques: Reinhard (R), Macenko (M), SPCN (S), ACD (A), StainGAN (SG) and StainNet (ST) on D_B to create stain-normalized dataset $D_{SN,T}$. Here, $SN \in \{R, M, S, A, SG, ST\}$ denotes the SN technique and $T \in \{T1, T2, T3, T4, T5, \varnothing\}$ (The $\varnothing$ is reserved for SG and ST where T is not required) signifies the template used. For example, $D_{R,T1}$ refers to the dataset normalized using the Reinhard technique with Template $T1$. Each dataset was split into a training set D_{TR} and a test set D_{TS} in an 80%-20% ratio (see Table 1 for the train test split).

We conducted Stratified Five-fold Cross-validation (SFFCV) on the training set D_{TR} by dividing it into five subsets, using one subset for validation and the remaining subsets for training. With SFFCV, we can compute the

Magnification	Benign	Malignant	Total
40x	625	1,370	1,995
100x	644	1,437	2,081
200x	623	1,390	2,013
400x	588	1,232	1,820
Total	2,480	5,429	7,909

Table 2. The BreaKHis image distribution by two major classes and four magnifications.

Architecture	Characteristic	FLOPs (B)	Parameters (M)
EfficientNet-B0 (EB0)[48]	Compound scaling	0.39	5.3
EfficientNet-V2-B0(EB0V2)[49]	Progressive learning	0.72	7.1
EfficientNet-V2-B0-21k (EB0V2-21k)[49]	Progressive learning	0.72	7.1
ResNet-V1-50 (RN1)[50]	Residual learning	4.1	25.6
ResNet-V2-50 (RN2)[51]	Identity mapping	4.1	25.6
MobileNet-V1 (MB1)[52]	Depth-wise separable convolutions	0.6	4.2
MobileNet-V2 (MB2)[53]	Inverted residuals and linear bottlenecks	0.3	3.4

Table 3. Description of the seven pre-trained CNNs in terms of their characteristics, number of FLOPs, and number of parameters.

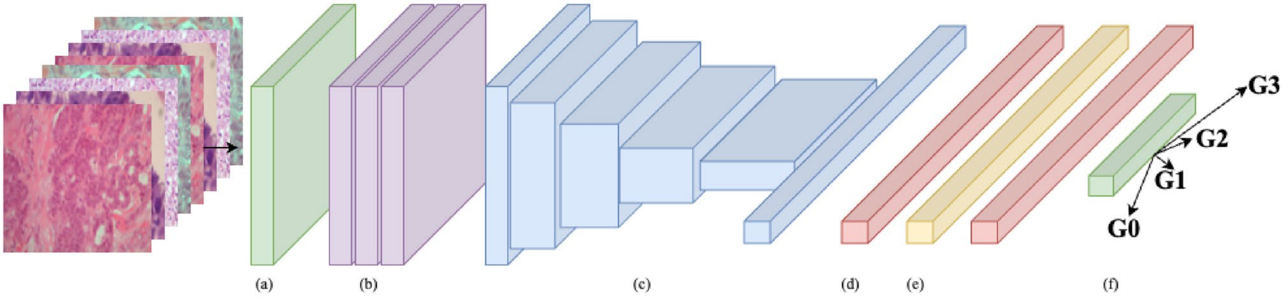

Figure 5. The structure of the model: (**a**) input layer, (**b**) augmentation layers, (**c**) feature extractor (non-trainable), (**d**) dropout layer, (**e**) dense layer (trainable), and (**f**) output prediction layer (trainable).

mean μ and standard deviation σ from results obtained from each subset for model stability evaluation (based on σ) and hyperparameters optimization. This process helps to minimize result variability, promote model stability, and provide a comprehensive performance evaluation across the base dataset D_B. After SFFCV, we retrained our models with the whole training set D_{TR} and tested on the testing set D_{TS} to obtain our baseline test result. Then, we repeated this procedure the stain-normalized training sets $D_{TR} \in D_{SN,T}$ and tested on the stain-normalized testing sets $D_{TS} \in D_{SN,T}$ to investigate the performance of CNN models trained with different stain-normalized datasets (see Algorithm *8*).

Before model training, we generated batches of pre-processed image data from each dataset with different image pre-processing functions (see Table 5). We also applied the class-weighting algorithm to address imbalanced classes in each dataset, ensuring the model converges for the minor classes in minimizing loss[70]. Equation (26) below describes the class-weighting algorithm.

$$ClassWeight = \frac{N}{N_c \times N_{sc}} \tag{26}$$

where N = number of images of all classes, N_c = number of classes and N_{sc} = number of images per class.

For the model implementation, we adhered to the approach outlined in Voon et al.[56]. We utilized seven pre-trained CNNs (see Table 3) from ImageNet[71] and ImageNet-21k[72] as feature extractors. Each model is composed of an input layer, augmentation layers, a feature *extractor* denoted as f_θ with model parameter θ, and a classifier denoted as $C(\cdot|W)$ with weight matrix $W \in \mathbb{R}^{d\times c}$. Our model structure is illustrated in Fig. 5. The classifier $C(\cdot|W)$ includes of two dropout layers and dense layers, with the final dense layer equipped with four neurons and a SoftMax activation function for classification (see Table 4). We kept the parameter θ in the f_θ fixed and trained a new classifier $C(\cdot|W)$ on each training set D_{TR} by minimizing the weighted categorical cross-entropy loss, $WCCE_{loss}$ (see Eq. (27)) using the Adam Optimizer[73]. Subsequently, we tested each trained classifier on its

Block	Detail
0	Input layer, shape = (224, 224, 3)
1	Augmentation layers:
	Random flip layer, mode = horizontal and vertical
	Random rotation layer, factor = 0.2
	Random zoom layer, height factor = 0.2
2	Feature extractor f_θ
3	Dropout layer, rate = 0.5
	Dense layer, 256 neurons with ReLU function
	Dropout layer, rate = 0.4
	Dense layer, 4 neurons with SoftMax function for final prediction

Table 4. The structure of the model which follows the implementation of Voon et al.[56].

	Operation	Value
Pre-processing function	Rescale	1./255
	Resize	224 by 224 pixels
	Shuffle	true
	Seed	123
	Batch	16
Hyperparameter	Loss function	$WCCE_{loss}$
	Optimizer	Adam
	Learning rate	0.001
	Metric	accuracy
	Epochs	100

Table 5. Details of image pre-processing and hyperparameters for model compilation.

corresponding testing set D_{TS}. The optimal learning rate and the number of epochs for model training were determined through SFFCV (see Table 5).

$$WCCE_{loss} = -w_j * log\left(\frac{e^{S_p}}{\sum_j^c e^{s_j}}\right) \tag{27}$$

where w_j = classes weights, $S_p =$ positive output score and $S_j =$ other classes output scores.

We primarily utilized the Balanced Accuracy (BAC) score as the evaluation metric for assessing model performance. The BAC, which calculates the average recall of each class, is computed using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The following mathematical expression defines the BAC:

$$BAC = \frac{1}{|N_c|}\sum_{i=1}^{|N_c|}\frac{TP_i}{TP_i + FN_i} \tag{28}$$

Results and discussion

Results of stratified five-fold cross-validation

Table 6 presents the cross-validation and test outcomes of the seven models trained on the base dataset D_B. Please note, the test result forms the baseline for subsequent comparisons. Interestingly, all models secured high BAC scores (> 0.9) in the base test set $D_{TS} \in D_B$. Among all models, the EB0V2-21k and MB1 models achieve the highest BAC score (0.9524). For the validation result, we observe that the EB0V2-21k model achieves the highest BAC with relatively high stability ($\mu = 0.9666$, $\sigma = 0.0185$). Generally, all models show low result variability. In other words, the models can generalize well across different subsets in D_{TR}.

Results of conventional stain normalization techniques

Figure 6, derived from Supplementary Tables 2–5, depicts the mean test BAC scores of seven models trained with datasets normalized using Reinhard, Macenko, SPCN, and ACD techniques across T. Our results underscore that the ACD technique yielded the highest average BAC score (0.905) across T, succeeded by Macenko (0.8835), SPCN (0.8567), and Reinhard (0.8407) techniques. Nonetheless, none of the techniques managed to surpass

```
Input: train set from the FBCG dataset, $D_{TR} \in D_B$, class weights $w$
Output: validation results, $val\_result$
for $K, (K - 1)$ in $SFFCV$ ($D_{TR}$) do
    for $epoch \in epochs$ do
        for $X, y$ batch $\in D_{(K - 1)}$ do
            $\hat{y} = C(X|W, w)$
            Compute $loss = WCCE_{loss}$ ($\hat{y}, y$)
            Compute the gradient, $\nabla_{loss}$ of the $W$ with respect to the $loss$
            Update $W \leftarrow Adam$ ($\nabla_{loss}, W$)
        end for
    end for
end for
$\hat{y} = C(D_k|W)$
$val\_result = BAC$ ($\hat{y}, y$)
```

Algorithm 7. SFFCV Model Training and Validation Loop

```
Input: train set, $D_{TR}$, test set, $D_{TS}$, and class weights $w$
Output: test results, $test\_result$
for $epoch \in epochs$ do
    for $X, y$ $batch \in D_{TR}$ do
        $\hat{y} = C(X|W, w)$
        Compute $loss = WCCE_{loss}$ ($\hat{y}, y$)
        Compute gradient, $\nabla_{loss}$ of the $W$ with respect to the $loss$
        Update $W \leftarrow Adam$ ($\nabla_{loss}, W$)
    end for
end for
$\hat{y} = C(D_{TS}|W)$
$test\_result = BAC$ ($\hat{y}, y$)
```

Algorithm 8. Model Training and Test Loop

Model	SFFCV ($\mu \pm \sigma$)	Test
EB0	0.9303 ± 0.0322	0.9518
EB0V2	0.9076 ± 0.0398	0.9024
EB0V2-21k	**0.9666 ± 0.0185**	**0.9524**
RN1	0.9253 ± 0.0310	0.9239
RN2	0.9346 ± 0.0156	0.9198
MB1	0.9518 ± 0.0232	**0.9524**
MB2	0.9362 ± 0.0322	0.9128
$\mu \pm \sigma$	**0.9361 ± 0.0189**	0.9308 ± 0.0211

Table 6. Cross-Validation and test BACs of seven models trained in D_B.The bolded values represent the highest score in each section.

the baseline result (0.9308). Among T, $T5$ yields the highest average BAC scores with Reinhard, Macenko, and SPCN techniques, whereas $T1$ attains the highest BAC using the ACD technique. $T5$ consistently achieves good results across different SN techniques. The superior performance of $T5$ may be attributed to the consideration of the dominant color in the target images. In histopathological images, the dominant color often corresponds to the stain used, which carries crucial information for classification tasks. By effectively capturing the dominant

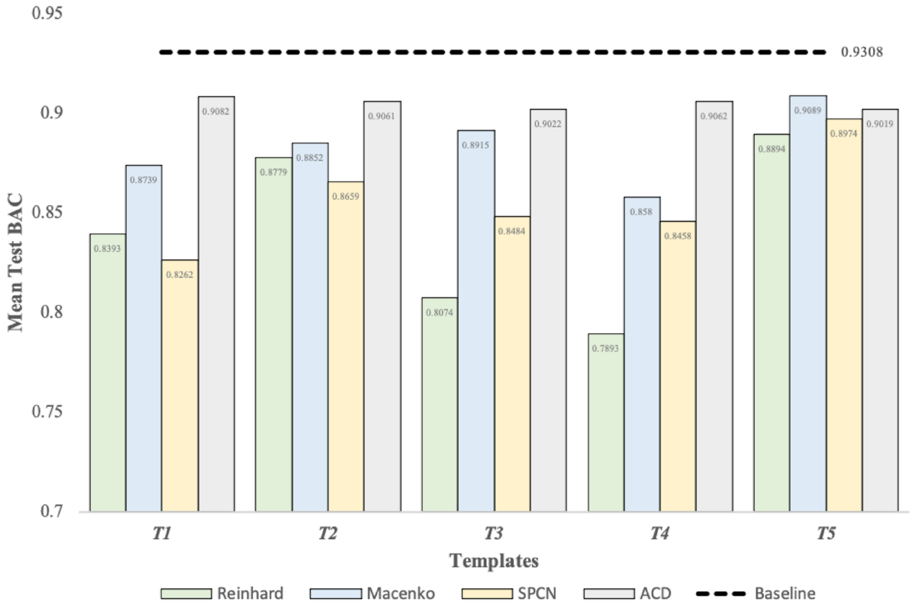

Figure 6. The mean test BAC scores of the seven models across *T* with different conventional SN techniques from Supplementary Tables 1, 2, 3, and 4. The ACD technique tops other techniques across all templates but failed to outperform the baseline result.

color, *T*5 can guide the SN process to better preserve or standardize this critical information, leading to improved classification performance.

Among conventional SN techniques, we noted that template selection minimally impacts the ACD technique due to its small σ (refer to Supplementary Table 4). In contrast, the Reinhard, Macenko and SPCN techniques are more heavily affected by the template selection. Hence, we propose that judicious template selection is crucial for Reinhard, Macenko, and SPCN techniques. Additionally, we suggest using the ACD technique for SN over other techniques if a conventional SN technique is required in the image pre-processing pipeline.

Results of deep learning-based stain normalization techniques

Figure 7, derived from Table 6 and Supplementary Table 5, depicts the test BAC scores of seven models trained with StainGAN-normalized, StainNet-normalized, and non-normalized datasets. We noted a high similarity in the performance of models trained with StainGAN-normalized and StainNet-normalized datasets, aligning with the findings by Kang et al.[47]. Nonetheless, models trained with the StainGAN-normalized dataset exhibited marginally higher mean test BAC scores (0.9196) than those trained with the StainNet-normalized dataset (0.9192). Additionally, our findings highlight that deep learning-based SN techniques failed to outperform the baseline result. Therefore, our results underscore the importance of context-specific application of these techniques and suggests that they may not universally lead to improved performance in every scenario.

Evaluation of the effectiveness of stain normalization in the idc grading task

In this section, we assessed the efficacy of SN in IDC grading using the FBCG dataset. Figure 8 illustrates the mean test BAC scores of the seven models trained in six different stain-normalized and the non-normalized datasets. Our results underscore that models trained with StainGAN-normalized images surpass those trained with other stain-normalized images. Hence, we compared the test mean BAC score between models trained with the StainGAN-normalized dataset and models trained with the non-normalized dataset. The results of the *t*-test indicated that the mean BAC score was statistically insignificant between models trained with the StainGAN-normalized FBCG dataset ($\mu = 0.9196$, $\sigma = 0.0188$) and models trained with the non-normalized dataset ($\mu = 0.9308$, $\sigma = 0.0211$), $p = 0.11$. The *p*-value indicates that the probability of obtaining the results is 11% by chance. Since the *p*-value of 0.11, higher than the significance level, $\alpha = 0.05$, suggests the difference in mean BAC scores between models trained with the StainGAN-normalized dataset and models trained with the non-normalized dataset is statistically insignificant. Consequently, we did not dismiss the null hypothesis, suggesting no significant difference in the performance of stain-normalized and non-normalized datasets for IDC grading tasks.

Furthermore, it is possible that SN techniques strip distinct color features[67] from IDC images, leading to poorer model performance. Our findings oppose the presumption that SN is essential to accomplish good performance in histopathological classification tasks, aligning with other similar studies[26–30]. Therefore, we suggest that future studies should conduct ablation studies with the employed dataset regarding the effectiveness of SN in their applications. Despite the ineffectiveness of SN in our IDC grading task, we acknowledge its contribution as evidenced by its benefits in other studies[31–35]. In response to the claim that SN may eliminate color features in IDC images, future studies could explore the influence of these color features on the generalizability of the CNN.

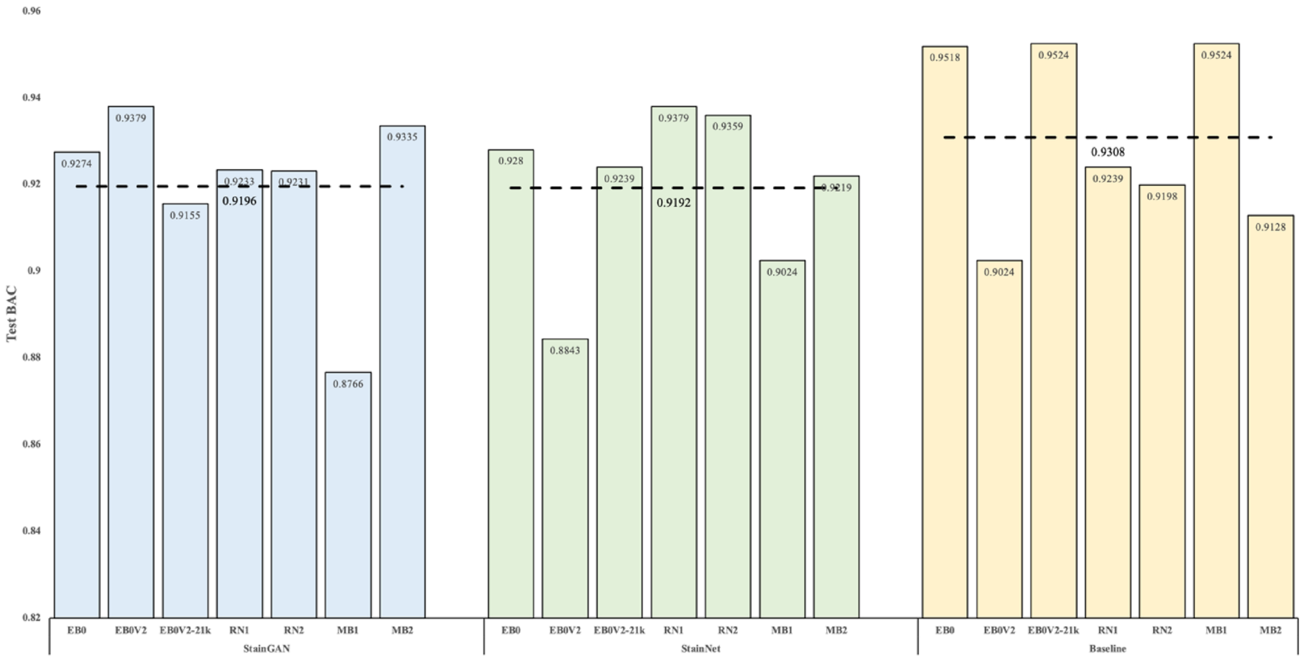

Figure 7. The test BAC scores of seven models trained with StainGAN-normalized, StainNet-normalized, and non-normalized datasets. Although the results are comparable among the deep learning-based SN techniques, the mean BAC scores of the seven models trained in the StainGAN-normalized dataset achieve slightly higher than models trained in the StainNet-normalized dataset but lower than the baseline result.

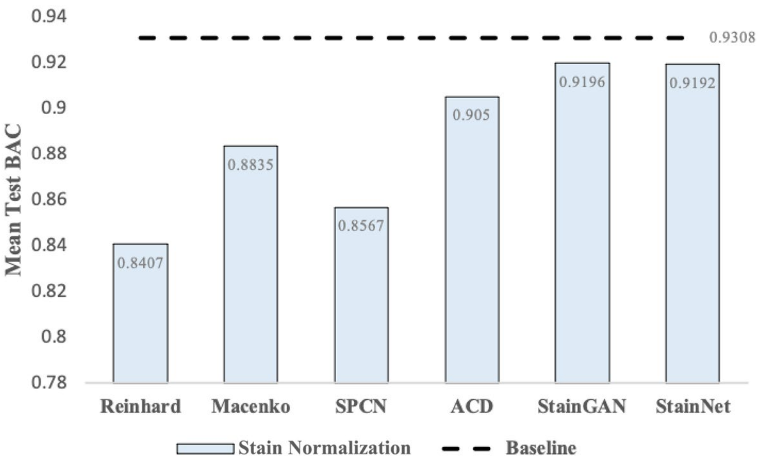

Figure 8. The mean test BAC scores of the seven models trained in six different stain-normalized and the non-normalized FBCG datasets. Among the six SN techniques, the StainGAN technique outperforms other SN techniques. However, the baseline result tops the best SN results by 0.0112 score.

In summary, the impact of SN on recent breast cancer histopathological studies has been the subject of debate. Our study aimed to elucidate this matter by scrutinizing the efficacy of SN techniques in breast cancer histopathological classification tasks, particularly in IDC grading, using CNNs. We selected six conventional and deep learning-based SN techniques to evaluate their effectiveness, along with seven pre-trained CNNs from ImageNet and ImageNet-21k as feature extractors. Our findings revealed that the impact of SN on this task was not statistically significant. Consequently, we did not reject the null hypothesis, suggesting that there was no substantial difference in effectiveness between stain-normalized and non-normalized datasets for IDC grading tasks. This outcome challenges the prevailing assumption that SN invariably enhances classification outcomes, thereby contributing a nuanced perspective to the discourse on the role of SN in breast cancer histopathological studies.

Limitations of study

The scope and limitations of our study focused to investigating the effectiveness of SN on IDC grading using only the FBCG dataset. Future work will incorporate other IDC grading datasets, such as DataBiox[74] and

PathoIDCG[55]. Additionally, our study did not account for potential variations in staining protocols across different centers. This is a significant consideration, since the staining process can greatly influence the color and intensity of histopathological images, which in turn can impact the performance of the model. While our findings underscore the impact of SN on IDC grading, they may not extend to scenarios where training and testing data come from separate centers. This limitation will be addressed in future work.

We utilized six different SN techniques in this study and plan to incorporate additional techniques[39,45,61,75,76] in future research. Subsequently, we selected five templates from the PCam train set to accommodate the Camelyon16 pre-trained StainGAN and StainNet. These templates were chosen as the results of applying three different similarity functions: (1) Cosine Similarity (SIM_C), (2) Mean Square Error (MSE), and (3) the Structural Similarity Index (SSIM), along with considering the most dominant color of the average image and the target images. The selection process aimed to identify templates that closely resemble the stain distributions in the target dataset. By using different similarity metrics, we were able to ensure that each template provided a unique perspective on the target data. Nonetheless, the five templates selected may not fully represent the color characteristics of the target dataset. This selection process has an empirical aspect, as there is no one-size-fits-all rule for template selection in style transfer.

For the model implementation, we only selected seven pre-trained CNNs for evaluations based on the implementation of Voon et al.[56]. We omitted other state-of-the-art CNNs[77–79] from our study but reserved them for future work. This study focused on the effectiveness of SN in the application; thus, we disregarded advanced model optimizations such as model fine-tuning and hyperparameter tuning.

Challenges of study

We encountered two significant challenges during the experimentation: (1) data imbalance and (2) model overfitting. An imbalanced dataset may inject bias into the CNN, causing the CNN to favor the majority class. Hence, we implemented the class-weighting algorithm that assigned higher weights to minority classes to increase the penalty. Given the relatively small size of our FBCG dataset compared to other breast cancer-related datasets, we noted a risk of model overfitting with complex CNN architectures. To mitigate this, we incorporated augmentation layers into our model for enhanced data diversity and added two dropout layers in our classifier to randomly nullify input units, thereby preventing overfitting during training.

Conclusion

In this study, we set out to address the question of the effectiveness of Stain Normalization (SN) in the task of Invasive Ductal Carcinoma (IDC) grading. To accomplish this, we utilized seven pre-trained Convolutional Neural Network (CNN) models as feature extractors to classify the FBCG dataset into four IDC grades. The FBCG dataset was stain-normalized using six techniques: Reinhard, Macenko, SPCN, ACD, StainGAN, and StainNet. For the conventional SN techniques, we selected five templates to investigate their impacts on each method. We conducted a comparative analysis of models trained with and without SN to understand the impact of SN on the classification results. Our findings revealed a p-value of 0.11 when comparing the test mean Balanced Accuracy (BAC) score of models trained with StainGAN-normalized (best-performing SN technique) images and non-normalized images. This indicates that there is no statistically significant difference in the effectiveness of stain-normalized and non-normalized datasets for IDC grading tasks. Contrary to common belief, our study suggests that SN may not be as crucial for histopathological classification tasks as previously thought. However, if SN is required in the image pre-processing pipeline, we recommend StainGAN, StainNet, and ACD techniques due to their relative performance in stain-normalizing images. Looking forward, in addition to extending our future work with the consideration mentioned in Sect. 4.5, we plan to examine the generalizability of the CNN model with respect to color features in IDC. Additionally, we aim to explore the inconsistent effects of SN on different breast cancer histopathological classification tasks.

Data availability

The origin datasets combined for the current study are available in the Four Breast Cancer Grades (FBCG) Dataset, https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/, and breast carcinoma histological images from the Department of Pathology, https://zenodo.org/record/834910#.WXhxt4jrPcs. Should there be any inquiries regarding the employed datasets, please contact the corresponding author, Dr. Hum Yan Chai (humyc@utar.edu.my) for further information and clarification.

Received: 28 February 2023; Accepted: 2 November 2023
Published online: 22 November 2023

References

1. Sharma, G. N., Dave, R., Sanadya, J., Sharma, P. & Sharma, K. K. Various types and management of breast cancer: An overview. *J. Adv. Pharm. Technol. Res.* **1**, 109–126 (2010).
2. Henson, D. E., Ries, L., Freedman, L. S. & Carriaga, M. Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. The basis for a prognostic index. *Cancer* **68**, 2142–2149 (1991).
3. Frkovic-Grazio, S. & Bracko, M. Long term prognostic value of Nottingham histological grade and its components in early (pT1n0m0) breast carcinoma. *J. Clin. Pathol.* **55**, 88–92 (2002).
4. Schwartz, A. M., Henson, D. E., Chen, D. & Rajamarthandan, S. Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumor size: A study of 161 708 cases of breast cancer from the SEER program. *Arch. Pathol. Lab. Med.* **138**, 1048–1052 (2014).
5. Rakha, E. A. *et al.* Breast cancer prognostic classification in the molecular era: The role of histological grade. *Breast Cancer Res.* **12**, 2607 (2010).

6. Johns Hopkins University. *Staging and Grade - Breast Pathology* (Johns Hopkins Pathology, 2021). https://pathology.jhu.edu/breast/staging-grade/.
7. He, L., Long, L. R., Antani, S. & Thoma, G. R. Histology image analysis for carcinoma detection and grading. *Comput. Methods Programs Biomed.* **107**, 538–556 (2012).
8. Elmore, J. G. *et al.* Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015).
9. Priego-Torres, B. M., Sanchez-Morillo, D., Fernandez-Granero, M. A. & Garcia-Rojo, M. Automatic segmentation of whole-slide H&E-stained breast histopathology images using a deep convolutional neural network architecture. *Expert Syst. Appl.* **151**, 113387 (2020).
10. Dalle, J.-R., Leow, W. K., Racoceanu, D., Tutac, A. E. & Putti, T. C. Automatic breast cancer grading of histopathological images. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2008**, 3052–3055 (2008).
11. Doyle, S., Agner, S., Madabhushi, A., Feldman, M. & Tomaszewski, J. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI* 496–499. https://doi.org/10.1109/ISBI.2008.4541041 (2008).
12. Naik, S. *et al.* Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI* 284–287. https://doi.org/10.1109/ISBI.2008.4540988 (2008).
13. Basavanhally, A. *et al.* Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides. *IEEE Trans. Biomed. Eng.* **60**, 2089–2099 (2013).
14. Dimitropoulos, K. *et al.* Grading of invasive breast carcinoma through Grassmannian VLAD encoding. *PLoS ONE* **12**, e0185110 (2017).
15. Wan, T., Cao, J., Chen, J. & Qin, Z. Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. *Neurocomputing* **229**, 34–44 (2017).
16. Li, L. *et al.* Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. *Multimed. Tools Appl.* **79**, 14509–14528 (2020).
17. Yan, R. *et al.* NANet: Nuclei-aware network for grading of breast cancer in HE stained pathological images. *Proceedings—2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020* 865–870. https://doi.org/10.1109/BIBM49941.2020.9313329 (2020).
18. Senousy, Z., Abdelsamea, M. M., Mohamed, M. M. & Gaber, M. M. 3E-net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images. *Entropy* **23**, 620 (2021).
19. Abdelli, A., Saouli, R., Djemal, K. & Youkana, I. Combined datasets for breast cancer grading based on multi-CNN architectures. *2020 10th International Conference on Image Processing Theory, Tools, and Applications, IPTA 2020*. https://doi.org/10.1109/IPTA50016.2020.9286653 (2020).
20. Zavareh, P. H., Safayari, A. & Bolhasani, H. *BCNet: A Deep Convolutional Neural Network for Breast Cancer Grading*. http://arxiv.org/abs/2107.05037 (2021).
21. Koo, J. C. *et al.* Non-annotated renal histopathological image analysis with deep ensemble learning. *Quant. Imaging Med. Surg.* **13**, 5902–5920 (2023).
22. Yong, M. P. *et al.* Histopathological gastric cancer detection on GasHisSDB dataset using deep ensemble learning. *Diagnostics* **13**, 1793 (2023).
23. Mccann, M. T. Tools for automated histology image analysis. *Carnegie Melon University Thesis Report* (2015).
24. McCann, M. T., Ozolek, J. A., Castro, C. A., Parvin, B. & Kovačević, J. Automated histology analysis: Opportunities for signal processing. *IEEE Signal Process. Mag.* **32**, 72–87. https://doi.org/10.1109/MSP.2014.2346443 (2015).
25. Ghaznavi, F., Evans, A., Madabhushi, A. & Feldman, M. Digital imaging in pathology: Whole-slide imaging and beyond. *Annu. Rev. Pathol.* **8**, 120902. https://doi.org/10.1146/annurev-pathol-011811-120902 (2013).
26. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).
27. Gupta, V., Singh, A., Sharma, K. & Bhavsar, A. Automated classification for breast cancer histopathology images: Is stain normalization important? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10550 LNCS**, 160–169 (2017).
28. Kumar, A. *et al.* Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Inf. Sci. (N Y)* **508**, 405–421 (2020).
29. Hameed, Z., Zahia, S., Garcia-Zapirain, B., Aguirre, J. J. & Vanegas, A. M. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors* **20**, 4373 (2020).
30. Hameed, Z., Garcia-Zapirain, B., Aguirre, J. J. & Isaza-Ruget, M. A. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. *Sci. Rep.* **12**, 1–21 (2022).
31. Nawaz, W., Ahmed, S., Tahir, A. & Khan, H. A. Classification Of Breast Cancer Histology Images Using ALEXNET. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10882 LNCS**, 869–876 (2018).
32. Munien, C. & Viriri, S. Classification of hematoxylin and eosin-stained breast cancer histology microscopy images using transfer learning with EfficientNets. *Comput. Intell. Neurosci.* **2021**, 1–17 (2021).
33. Salvi, M., Molinari, F., Acharya, U. R., Molinaro, L. & Meiburger, K. M. Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification. *Comput. Methods Programs Biomed Update* **1**, 100004 (2021).
34. Alkassar, S., Jebur, B. A., Abdullah, M. A. M., Al-Khalidy, J. H. & Chambers, J. A. Going deeper: Magnification-invariant approach for breast cancer classification using histopathological images. *IET Comput. Vis.* **15**, 151–164 (2021).
35. Shahidi, F., Daud, S. M., Abas, H., Ahmad, N. A. & Maarop, N. Breast cancer classification using deep learning approaches and histopathology image: A comparison study. *IEEE Access* **8**, 187531–187552 (2020).
36. Ehteshami Bejnordi, B. *et al.* stain specific standardization of whole-slide histopathological images. *IEEE Trans. Med. Imaging* **35**, 404–415 (2016).
37. Bautista, P., Hashimoto, N. & Yagi, Y. Color standardization in whole slide imaging using a color calibration slide. *J. Pathol. Inform.* **5**, 4 (2014).
38. Veta, M., Pluim, J. P. W., van Diest, P. J. & Viergever, M. A. Breast cancer histopathology image analysis: A review. *IEEE Trans. Biomed. Eng.* **61**, 1400–1411 (2014).
39. Khan, A. M., Rajpoot, N., Treanor, D. & Magee, D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans. Biomed. Eng.* **61**, 1729–1738 (2014).
40. Reinhard, E., Ashikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph Appl.* **21**, 34–41 (2001).
41. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. In *IEEE International Symposium on Biomedical Imaging* 1107–1110 (IEEE, 2009).
42. Vahadane, A. *et al.* Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging* **35**, 1962–1971 (2016).

43. Zheng, Y. *et al.* Adaptive color deconvolution for histological WSI normalization. *Comput. Methods Programs Biomed.* **170**, 107–120 (2019).
44. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation Equivariant CNNs for Digital Pathology. *arXiv preprint arXiv* **11071 LNCS**, 210–218 (2018).
45. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
46. Shaban, M. T., Baur, C., Navab, N. & Albarqouni, S. Staingan: Stain style transfer for digital histological images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* vols 2019-April 953–956 (IEEE, 2019).
47. Kang, H. *et al.* StainNet: A fast and robust stain normalization network. *Front. Med.* **8**, 746307 (2021).
48. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning, ICML 2019* vols 2019-June 10691–10700 (International Machine Learning Society (IMLS), 2019).
49. Tan, M. & Le, Q. V. *EfficientNetV2: Smaller Models and Faster Training* (2021).
50. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-Decem**, 770–778 (2015).
51. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9908 LNCS**, 630–645 (2016).
52. Howard, A. G. *et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* (2017).
53. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4510–4520 (IEEE, 2018). https://doi.org/10.1109/CVPR.2018.00474.
54. Pan, X. *et al.* Multi-task deep learning for fine-grained classification/grading in breast cancer histopathological images. *Stud. Comput. Intell.* **810**, 85–95 (2020).
55. Yan, R. *et al.* Nuclei-guided network for breast cancer grading in HE-stained pathological images. *Sensors* **22**, 4061 (2022).
56. Voon, W. *et al.* Performance analysis of seven Convolutional Neural Networks (CNNs) with transfer learning for Invasive Ductal Carcinoma (IDC) grading in breast histopathological images. *Sci. Rep.* **12**, 19200 (2022).
57. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
58. Zheng, Y. *et al.* Stain standardization capsule for application-driven histopathological image normalization. *IEEE J. Biomed. Health Inform.* **25**, 337–347 (2021).
59. Zhou, N., Cai, D., Han, X. & Yao, J. Enhanced cycle-consistent generative adversarial network for color normalization of H&E stained images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11764 694–702 (2019).
60. Zanjani, F. G., Zinger, S., de With, P. H. N., Bejnordi, B. E. & van der Laak, J. Histopathology stain-color normalization using deep generative models. *Medical Imaging with Deep Learning* (2018).
61. Lei, G. *et al.* StainCNNs: An efficient stain feature learning method. *Neurocomputing* **406**, 267–273 (2020).
62. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October 2242–2251 (2017).
63. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **63**, 1455–1462 (2016).
64. Bándi, P. *et al.* 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 (2019).
65. Aresta, G. *et al.* BACH: Grand challenge on breast cancer histology images. *Med. Image Anal.* **56**, 122–139 (2019).
66. Salvi, M., Michielli, N. & Molinari, F. Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. *Comput. Methods Programs Biomed.* **193**, 105506 (2020).
67. Bianconi, F. *et al.* Evaluation of colour pre-processing on patch-based classification of H&E-stained images. *Eur. Congr. Digit. Pathol.* **11435**, 56–64 (2019).
68. Roy, S., Kumar Jain, A., Lal, S. & Kini, J. A study about color normalization methods for histopathology images. *Micron* **114**, 42–61 (2018).
69. Zioga, C. *et al.* Breast carcinoma histological images from the Department of Pathology, 'Agios Pavlos' General Hospital of Thessaloniki, Greece, Greece. ZENODO https://doi.org/10.5281/ZENODO.834910 (2017).
70. Analytics Vidhya. How to dealing with imbalanced classes in machine learning. *Analytics Vidhya* https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/ (2020).
71. Deng, J. *et al. ImageNet: A Large-Scale Hierarchical Image Database* 248–255 https://doi.org/10.1109/CVPR.2009.5206848. (2010).
72. Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. *ImageNet-21K Pretraining for the Masses* (2021).
73. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015* (International Conference on Learning Representations, ICLR, 2014). https://doi.org/10.48550/arxiv.1412.6980.
74. Bolhasani, H., Amjadi, E., Tabatabaeian, M. & Jassbi, S. J. A histopathological image dataset for grading breast invasive ductal carcinomas. *Inform. Med. Unlocked* **19**, 100341 (2020).
75. Lee, C.-C., Kuo, P.-T. P. & Peng, C.-H. H&E stain normalization using U-net. In *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)* 29–32 (IEEE, 2022). https://doi.org/10.1109/BIBE55377.2022.00014.
76. Pérez-Bueno, F. *et al.* Bayesian K-SVD for H and E blind color deconvolution. Applications to stain normalization, data augmentation and cancer classification. *Computerized Medical Imaging and Graphics* **97**, 102048 (2022).
77. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* vols 2017-January 2261–2269 (Institute of Electrical and Electronics Engineers Inc., 2016).
78. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 201—Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2014). https://doi.org/10.48550/arxiv.1409.1556.
79. Liu, Z. *et al.* A ConvNet for the 2020s. *arXiv preprint arXiv* 11966–11976 https://doi.org/10.48550/arxiv.2201.03545 (2022).

Author contributions

In this study, W.V was responsible for conducting a comparative analysis and drafting the main manuscript. Y.C.H designed the methodology, produced the figures, and oversaw the research. Y.K.T, W.S.Y, and H.N provided assistance with result interpretation. H.M validated the findings, while N.G and K.W.L explored the data.

Funding

The authors would like to gratefully acknowledge the support of Universiti Tunku Abdul Rahman Research Fund (Ref: IPSR/RMC/UTARRF/2022-C1/H01).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46619-6.

Correspondence and requests for materials should be addressed to Y.C.H.