Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.7045

Quantifying Color and Distortion Biases in the NCT-CRC-HE-100K Histopathology Dataset

Gabriel Arquelau Pimenta Rodrigues [**University of Brasilia** | ***gabriel.arquelau@redes.unb.br***]
André Luiz Marques Serrano [**University of Brasilia** | ***andrelms@unb.br***]
Geraldo Pereira Rocha Filho [**State University of Southwest Bahia** | ***geraldo.rocha@uesb.edu.br***]
Rodrigo Bonacin [**Renato Archer Information Technology Center** | ***rodrigo.bonacin@cti.gov.br***]
Vinícius Pereira Gonçalves [**University of Brasilia** | ***vpgvinicius@unb.br***]
Muttukrishnan Rajarajan [**City St George's, University of London** | ***r.muttukrishnan@city.ac.uk***]
Rodolfo Ipolito Meneguette [**University of São Paulo** | ***meneguette@icmc.usp.br***]

Department of Electrical Engineering (ENE), Faculty of Technology, University of Brasília (UnB), Brasília, DF, 70910-900.

Received: 29 November 2025 • **Accepted:** 10 April 2026 • **Published:** 07 May 2026

Abstract Colorectal cancer (CRC) represents a persistent challenge for healthcare systems, and the development of reliable deep learning systems for histopathology depends on unbiased datasets. The widely used NCT-CRC-HE-100K dataset has been shown to contain color inconsistencies, distortion artifacts, and corrupted patches, yet prior analyses offered only limited quantitative evidence. In this work, we extend these observations by evaluating color signatures, stain-normalization behavior, and class-dependent image quality variations. We compare classical and deep learning based stain normalization methods to identify their impact on image quality metrics and potential reduction of class-specific biases in computational pathology. Our results show that while normalization reduces color-based class distinguishability, none of the evaluated methods completely eliminate tissue-specific color signatures. Additionally, this work demonstrates that distortion artifacts disproportionately affect one class in the dataset, introducing technical biases unrelated to morphology. Also, a CNN classifier trained on each normalized dataset indicates that model performance is not significantly changed across the normalization methods, including the unnormalized dataset, despite reductions in color-based separability. Overall, our study provides quantitative evidence that color, saturation, and distortion persist across normalization techniques, emphasizing the need for caution when using NCT-CRC-HE-100K to assess histopathology models.

Keywords: Bias analysis, colorectal cancer, histopathology, stain normalization

1 Introduction

Cancer is a significant causes of mortality worldwide [Filho *et al*., 2025]. Compounding this issue, an increase is expected in the global cases of colorectal cancer from 1.9 million in 2019 to 3.2 million in 2040 [Xi and Xu, 2021]. Thus, early and accurate cancer diagnosis is fundamental to improving public health. To achieve this, advances in medical imaging and in deep learning techniques have facilitated the development of automated cancer detection [Jiang *et al*., 2023].

Evidence indicates that lifestyle factors, particularly diet, are major contributors to the incidence of colon cancer [Yu *et al*., 2022]. The consumption of ultra-processed foods (UPF), such as soft drinks and reconstituted meat, has been linked to an increased risk of gastrointestinal tract malignancies [Levy *et al*., 2024]. These dietary components, characterized by high levels of additives and unhealthy fats, may induce inflammation and alter gut microbiota. Processed foods are known to be harmful, and growing evidence suggests they may also contribute to cancer development, particularly in organs like the colon [Meine *et al*., 2024; Cordova *et al*., 2023; Wang *et al*., 2022b].

Furthermore, consumption of UPF has increased in the United States over the past two decades [Juul *et al*., 2022], and it is among the 20 most consumed foods in Brazil [Rodrigues *et al*., 2021]. This dietary trend is also observed among infants in low- and middle-income countries [Popkin and Laar, 2025], which emphasizes the need for improved colon cancer screening and early detection methods to address the potential rise in cases resulting from these lifestyle changes.

Given this alarming dietary change, there is a need for scalable diagnostic tools. Automated classification in medical diagnostics assists doctors by providing information to support their diagnostic reasoning. This leads to earlier and more accurate disease detection, enabling timely, targeted treatments that improve patient outcomes [Di Giammarco *et al*., 2024]. With greater accuracy, clinicians can reduce unnecessary tests and avoid misdiagnoses, enhancing patient safety and healthcare efficiency.

To achieve this level of automated diagnostic precision, large-scale, well-annotated datasets are fundamental. They are necessary for advancing algorithm development, facilitating reproducible research, and establishing performance benchmarks to meet the standards required for medical diagnostic applications. An example of such a dataset for colorectal cancer is the NCT-CRC-HE-100K dataset [Kather *et al*., 2018]. It has been used as a foundation for several studies

[Uddin *et al*., 2024; Pacal and Attallah, 2025].

This dataset, however, contains class-dependent color signatures, color oversaturation, and heterogeneous distortions, as observed by Ignatov and Malivenko [2024]. These biases may allow classifiers to exploit technical artifacts instead of morphology. A color bias enables classifiers to achieve high accuracy by learning staining artifacts rather than biological morphology. Furthermore, the dataset contains inconsistent, distorted, and corrupted images. Consequently, classification models may learn to rely on these technical artifacts as spurious correlates of tissue classes rather than on histopathological morphology.

In this work, we extend the observations of Ignatov and Malivenko [2024] with a quantitative analysis of the biases present in NCT-CRC-HE-100K-NONORM. We measure class-dependent color signatures, compare the behavior of stain-normalization methods, and quantify the prevalence and severity of image corruptions and distortion artifacts across the tissue types. This analysis identifies which classes are most affected by these issues, which may correlate with misclassification patterns in deep learning models trained on this dataset.

While prior studies have reported descriptive observations of color artifacts and normalization effects in histopathology datasets Ignatov and Malivenko [2024], this study provides a statistically validated analysis that quantifies the prevalence and severity of color bias and distortion artifacts and tests their dependence on different color normalization implementations. Additionally, we investigate whether distortion-related signals are sufficient to explain the observed bias by isolating high-frequency components associated with compression artifacts.

1.1 Contributions and limitations of this work

While Ignatov and Malivenko [2024] identified color inconsistencies, JPEG artifacts, and corrupted patches, their analysis did not fully quantify their prevalence, severity, or class-dependent structure, nor compare the behavior of different stain-normalization strategies. Thus, this work contributes a quantitative extension of the dataset analysis, focusing on measuring these artifacts, with inferential statistical testing to validate bias claims.

This work compares several stain-normalization approaches, including both classical and deep learning-based methods, and evaluates how they reshape the color distributions across tissue classes, highlighting cases in which normalization reduces bias and others where it fails to achieve this reduction. We also quantify the severity and prevalence of corrupted and distorted images, demonstrating that these distortions are not uniformly distributed and exhibit class-dependent patterns. Hence, this study extends previously descriptive observations into reproducible and interpretable evidence that clarifies how low-level image properties influence model predictions.

Our study, however, is limited to the NCT-CRC-HE-100K dataset and may not generalize to datasets with different staining or acquisition protocols. We focus on patch-level properties and do not analyze patient-level effects. Additionally, although we compare the performance of several normalization techniques, we do not propose new normalization algorithms or dataset corrections.

In this study, all stain normalization methods that require a reference use the same single reference image as the target, to which the color distribution of all other images is normalized, rather than multiple targets. Although consistent with standard implementations, this choice may not adequately represent the stain variability present across different tissue types, potentially limiting the normalization results.

Furthermore, although stratified sampling reduces computational costs while maintaining statistical properties, the selected subset may not represent rare artifacts in the full dataset. Consequently, our findings reflect statistically validated tendencies within the sampled subset, rather than exhaustive measurements of all potential distortions.

1.2 Structure of this work

The remainder of this work is structured as follows. Section 2 discusses related works. Section 3 presents the methodology adopted in this study, whilst Section 4 details the results obtained. Section 5 concludes the paper.

2 Literature review

Given the criticality of the disease, several works have proposed deep learning models to diagnose colorectal cancer (CRC), mainly because these Artificial Intelligence (AI) models have demonstrated to improve diagnostic accuracy in histopathological analysis [Merabet *et al*., 2025].

To advance research in this area and ensure the development of generalizable AI, large-scale, well-annotated histopathological image datasets are necessary for training, testing, and validating these models. For example, Borkowski *et al*. [2019] proposed the LC25000 dataset, comprising 25,000 color images across 5 classes of benign and malignant lung and colon tissues. Other datasets focused on CRC are Sirinukunwattana *et al*. [2017]; Barbano *et al*. [2021]; Rezaei *et al*. [2023]; Shi *et al*. [2023]; Mokhtari *et al*. [2023].

Another example of a CRC dataset is the NCT-CRC-HE-100K [Kather *et al*., 2018], which has been used as a foundation for several deep learning works. For instance, Li [2024] proposes a Swin-Transformer V2 model for colorectal cancer tissue classification, achieving a top-l accuracy of 96.0%. The approach incorporates self-supervised pre-training on tumor-related datasets and employs a progressive layer-wise distillation technique to transfer knowledge from a larger teacher model.

Furthermore, Qin *et al*. [2024] integrates a skip feedback connection structure into a U-Net framework and combines it with the Swin Transformer for enhanced feature extraction, aiming to improve the model's multi-level feature learning capabilities. Their algorithm enables end-to-end recognition of colorectal adenocarcinoma tissue images and achieves 95.8% accuracy on the NCT-CRC-HE-100K dataset. Additionally, Firildak *et al*. [2025] propose an architecture that combines ReFeatureBlock, depthwise convolution, and global average pooling to extract discriminative features while reducing computational complexity. The model achieves 99.19% accuracy

Table 1. Comparison between prior work and this study (✓: yes, ○: limited, ✗: no). Statistical validation refers to the use of formal inferential tests to support claims

Study	Histopathology	Quantify bias	Statistical validation	Compare color norm. tech.
Wang *et al*. [2022a]	✗	✓	✓	✗
Rinaldi *et al*. [2022]	✓	✗	✗	✓
Hoque *et al*. [2024]	✓	✗	✗	✓
Tellez *et al*. [2019]	✓	○	○	✓
Ignatov and Malivenko [2024]	✓	✓	○	✗
This work	✓	✓	✓	✓

on NCT-CRC-HE-100K.

These works, however, may be exploiting class-dependent biases present in the dataset, as Ignatov and Malivenko [2024] reveals limitations that challenge the validity of reported deep learning results in colorectal cancer histopathology using the NCT-CRC-HE-100K dataset. The authors demonstrate the presence of technical artifacts in the images that may introduce class biases, including inappropriate color normalization, inconsistent distortions, and corrupted tissue samples.

Their work also shows that models using only RGB color intensities achieve over 50% accuracy on this 9-class classification task, while color histogram features yield 82% accuracy without capturing any morphological information. A basic EfficientNet-B0 model achieves 97.7% accuracy, outperforming some of the previously proposed specialized architectures. These findings question whether reported high performances reflect genuine histological understanding or merely the learning of dataset-specific technical artifacts, urging careful reinterpretation of results obtained on this widely used benchmark. Our work extends the study of Ignatov and Malivenko [2024] to quantify these biases and compare them across several color normalization techniques.

These biases are also present in other datasets, such as in The Cancer Genome Atlas (TCGA) [Dehkharghanian *et al*., 2023]. This implies that the reported high performance models that use this dataset may not reflect genuine diagnostic capability but rather the model's exploitation of dataset-specific artifacts, which can evidence site-specific biases [Kheiri *et al*., 2025]. In fact, most visual datasets suffer from some kind of bias [Fabbrizzi *et al*., 2022].

One possible source of bias is staining variation in hematoxylin and eosin (H&E) images, which arises from differences in laboratory protocols, scanners, and chemical batches. Without normalization, deep learning models may learn to recognize specific staining patterns rather than pathological features. Because of this, it is important to apply color normalization in the images. Roy *et al*. [2018] compared some color normalization methods in histopathology images and concluded that the approach proposed by Vahadane *et al*. [2016] achieved the best results in the evaluated metrics. Similar results were obtained by Hoque *et al*. [2024].

Considering this, our work includes this color normalization approach in the comparison for the NCT-CRC-HE-NONORM dataset, along with several others. Distortion artifacts may also introduce bias in the dataset [Liu and He, 2024].

Moreover, HistoQC is an open-source tool developed by Janowczyk *et al*. [2019] for quality control of H&E-stained whole-slide images (WSI). It automatically detects technical artifacts, such as tissue folds, pen markings, blurring, and uneven staining, and provides quantitative metrics for brightness, contrast, color, and tissue coverage. Thus, HistoQC can also help identify potential class-specific biases in datasets that might otherwise lead models to learn technical artifacts instead of biological features. However, HistoQC cannot be applied in this work, as the available images are small, single-resolution TIF files rather than multi-resolution WSI compatible with OpenSlide.

Table 1 summarizes the main contributions of this study in comparison with related work.

3 Methodology

The methodological framework adopted in this study, illustrated in Figure 1, follows a sequential approach to data processing and analysis, which is further detailed in this section. Python version 3.12.12 is used throughout the work.

All experiments were conducted on a Linux-based system using a CPU-only configuration. The machine was equipped with an x86-64 processor with 2 CPU cores and 16 GB of RAM.

3.1 Dataset

The dataset[1] published by Kather *et al*. [2018] is used in this work. It comprises 100,000 non-overlapping H&E stained histopathological image patches extracted from CRC tissue samples, annotated across nine distinct tissues classes, namely adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM).

All histopathological TIF images in the dataset have a resolution of 224×224 pixels and a spatial resolution of 0.5 microns per pixel. The authors provide two distinct versions of the dataset: the NCT-CRC-HE-100K collection, with images normalized using the method proposed by Macenko *et al*. [2009]; and the NCT-CRC-HE-100K-NONORM collection, comprising the original images without color normalization.

In this work, we use the NCT-CRC-HE-100K-NONORM to compare different color normalization techniques and NCT-CRC-HE-100K for the other analyses. Table 2 summarizes which dataset was used in each analysis.

[1] `https://zenodo.org/records/1214456`

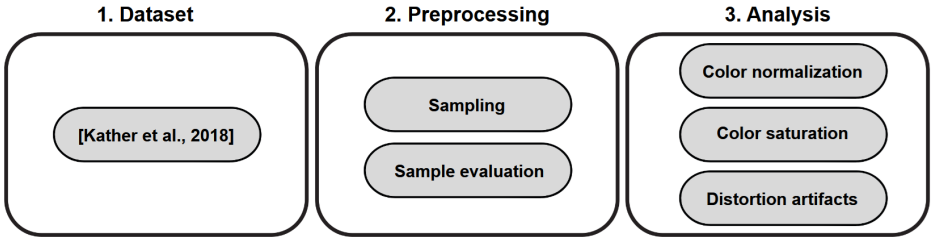


Figure 1. Diagram representing the steps of this work

Table 2. Dataset usage across analyses. NONORM refers to NCT-CRC-HE-100K-NONORM, while 100K refers to NCT-CRC-HE-100K.

Analysis	Sections	Dataset
Color normalization	3.3.1, 4.1	NONORM
Color saturation	3.3.2, 4.1.3	100K
Distortion artifacts	3.3.3, 4.2	100K

3.2 Preprocessing

The used datasets present practical challenges for iterative experimentation due to their substantial volume of 100,000 images. Therefore, a stratified sampling strategy is employed to create a representative subset, as described in this section.

3.2.1 Sampling

Due to the class imbalance, with different tissue types occurring at different frequencies, a stratified sampling approach is adopted to preserve these proportional distributions [Lohr, 2021].

$$\varepsilon = z \cdot \sqrt{\sum_{h=1}^{9} \left(\frac{N_h}{N}\right)^2 \cdot \frac{p(1-p)}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right)} \quad (1)$$

To determine the margin of error for the sampled dataset, we applied the Equation defined in 1, where N is the total population size ($N = 100,000$), n_h represents the number of samples allocated to class h, N_h is the population size of the class h, $p = 0.5$ is the expected proportion for maximum variability, $z = 1.96$ is adopted to the corresponding 95% confidence level, and ε is the achieved margin of error. Sampling was performed once using a fixed random seed of 42, and the same stratified subset was reused across all analyses and experiments.

For a total sample size of 13,500 images, this stratified approach yields a margin of error of approximately $\pm 0.91\%$. The resulting sample sizes per class are presented in Table 3.

3.2.2 Sample evaluation

A margin of error of $\pm 0.91\%$ provides a significant degree of statistical precision for the sample's representativeness. However, to further validate that the sampled dataset preserves the characteristics of the full population, we use the Kolmogorov-Smirnov (KS) test and Mann-Whitney U test.

The KS test evaluates whether two distributions differ by comparing their empirical cumulative distributions. The null hypothesis (H_0) states that there is no meaningful difference between the distributions of the sample and the reference group. Conversely, the Mann-Whitney U test evaluates whether two groups differ by comparing their ranked observations. The null hypothesis (H_0) states that there is no systematic difference in central tendency between the two groups.

In the context of sample evaluation, the KS test verifies whether the sample preserves the overall distributional shape of the group it represents. In contrast, the Mann-Whitney U test examines whether the sample and the original dataset differ in their central tendency.

Table 3. Samples sizes per class

Class	Original size	Sample size
ADI	10,407	1,405
BACK	10,566	1,426
DEB	11,512	1,554
LYM	11,557	1,560
MUC	8,896	1,202
MUS	13,536	1,827
NORM	8,763	1,183
STR	10,446	1,410
TUM	14,317	1,933
Total	100,000	13,500

Table 4. Statistical validation of sample representativeness

Feature	KS p-value	M-W p-value
Mean Intensity	0.954	0.900
Color STD	0.572	0.792
Entropy	0.947	0.623
Contrast	0.645	0.675
Color Range	0.946	0.855

The selected features were chosen because they reflect the image properties investigated in this work, namely color bias and distortion effects introduced by stain normalization.

As the p-values for both tests are above the 0.05 threshold for the 95% confidence level for all measured features, as shown in Table 4, the sample can be considered statistically representative of the full dataset with respect to the color and distortion-related properties analyzed in this study.

3.3 Analysis

For the dataset analysis, we focus on the problems identified by Ignatov and Malivenko [2024].

3.3.1 Color normalization

As observed by Ignatov and Malivenko [2024], even with the Macenko *et al*. [2009] color normalization, the NCT-CRC-HE-100K dataset still exhibits a color signature, which helps classification models achieve higher accuracy.

To quantify this tissue class separability, we use a logistic regression classification model based exclusively on intensity and color histogram. As it does not consider any biological structure, a good color normalization will reduce the accuracy of this model. We also measure the mean Wasserstein distance and the mean Maximum Mean Discrepancy (MMD) between the classes.

Each image is represented by a 40-dimensional color-only feature vector composed of (i) mean and standard deviation of RGB intensities (6 features); (ii) normalized RGB color histograms with 10 bins per channel (30 features); and (iii) mean and standard deviation of the hematoxylin and eosin channels obtained via RGB-to-HED conversion (4 features). No spatial or morphological descriptors are used.

Color-only classification accuracy is estimated using 5-fold stratified cross-validation with shuffled splits (random seed 42). Feature standardization is applied within each fold.

Complementarily, we use the mean Structural Similarity Index (SSIM) and the mean Peak Signal-to-Noise Ratio (PSNR) to assess image fidelity relative to the non-normalized corresponding image.

Using these metrics, we compare different color normalization techniques, encompassing both classical approaches, including the methods by Ruifrok [2001], Reinhard *et al*. [2002], Macenko *et al*. [2009], and Vahadane *et al*. [2016]; and deep learning based strategies, namely StainGAN [Shaban *et al*., 2019] and StainNet [Kang *et al*., 2021]. For a reproducible implementation of the classical methods, we utilize the TIAToolbox [Pocock *et al*., 2022], version 1.6.0.

3.3.2 Color saturation

Ignatov and Malivenko [2024] concluded that the DEB class exhibits variations in color saturation and channel-specific artifacts. These colorimetric disparities may introduce biases and potentially affect the performance of computational models by creating class-specific staining patterns. They also suggest that it may be a consequence of the color normalization process. Since it has been observed only in the DEB class, we restrict the color saturation analysis to this class.

To quantify these color channel variations across different normalization techniques, we measure the Blue Dominance (BD) of the image, which quantifies the relative prominence of blue staining, as defined per Equation 2.

$$\mathrm{BD} = \frac{\mu_B}{\mu_R + \mu_G + \epsilon} \tag{2}$$

Where μ_B represents the mean intensity of the blue channel, μ_R represents the mean intensity of the red channel, μ_G represents the mean intensity of the green channel, and $\epsilon = 1 \times 10^{-6}$ is a small constant added to prevent division by zero. All intensity values are normalized to the range [0,1] before calculation.

The theoretical range of the BD metric spans from 0 to infinity. As the denominator approaches zero in pure blue images, the metric approaches infinity, while completely blue-free images approach zero.

3.3.3 Distortion artifacts

It has also been observed that certain tissue classes in histopathological imaging datasets exhibit greater distortion than others, due to the presence of JPEG compression artifacts [Ignatov and Malivenko, 2024]. These quality disparities potentially introduce class-specific biases that could confound computational analysis.

To quantify these quality variations across tissue classes, we calculate the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) using the Python Image Quality (PIQ) library [Kastryulin *et al*., 2022], version 0.8.0. BRISQUE operates in the spatial domain to detect distortion artifacts without requiring a reference image [Mittal *et al*., 2012]. The higher the BRISQUE score, the poorer the quality of the assessed image, which may be due to blurring, noise, contrast changes, compression artifacts, or color distortions.

Image distortions identified by BRISQUE may be more noticeable in tissue regions with homogeneous texture, as they become more pronounced. In contrast, regions with more complex texture can mask these artifacts, making them harder to detect. To quantify this, we compute Total Variation (TV) as a measure of image homogeneity, as in Equation 3.

$$\mathrm{TV} = \sum_{i,j} \left(|\nabla_x I(i,j)| + |\nabla_y I(i,j)| \right) \tag{3}$$

The calculation requires computing horizontal ($\nabla_x I(i,j)$) and vertical ($\nabla_y I(i,j)$) gradients at each pixel position (i,j), which measure the rate of intensity change in their respective directions. Lower TV values indicate more homogeneous regions, while higher TV values correspond to textured areas with abundant structural details.

These metrics are calculated per class to identify those with a higher proportion of distortion artifacts.

4 Results

This section presents the results of the dataset analysis.

4.1 Color-based bias analysis

This section examines chromatic bias in the dataset, analyzing the impact of stain normalization and channel-specific saturation artifacts.

4.1.1 Normalization comparison

To mitigate class-specific color signatures in the dataset, several color normalization techniques are applied. Among the methods evaluated, Ruifrok, Vahadane, Macenko, and Reinhard require a target image that defines the desired color distribution. This reference image is not required for StainNet nor StainGAN.

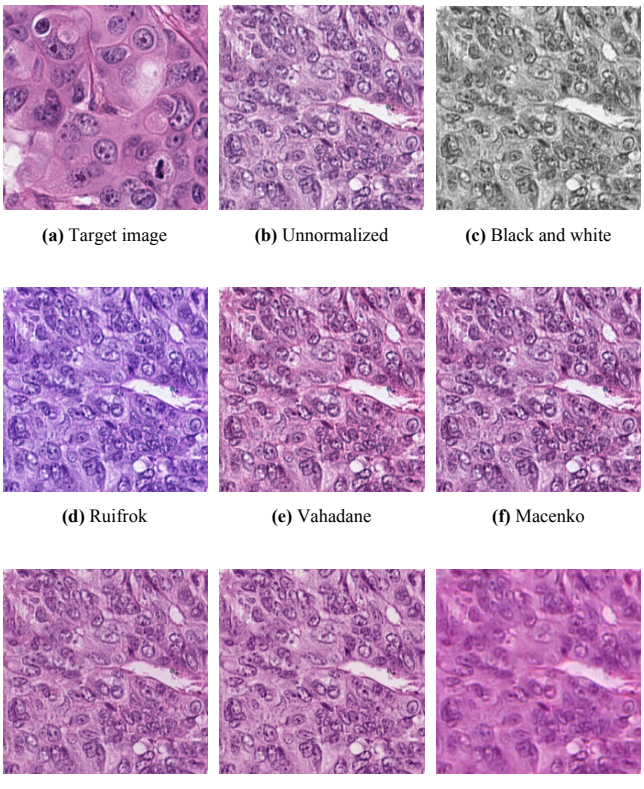
A single target image is selected to ensure consistency across all normalization techniques. The selected target image, `TUM-AAPIVHNA.tif`, shown in Figure 2a, exhibits bal-

Table 5. Comparison of color normalization techniques using class separability and image quality metrics

Method	Color Acc.	Acc. 95% CI	CLD	Mean Wass.	Mean MMD	Mean SSIM	Mean PSNR
Unnormalized	0.823 ± 0.009	(0.812-0.834)	A	5.200	0.057	-	-
B/W	0.604 ± 0.006	(0.595-0.609)	B	5.476	0.061	0.943	24.75
Ruifrok	0.813 ± 0.005	(0.803-0.816)	A	2.913	0.013	0.843	16.61
Vahadane	0.788 ± 0.010	(0.770-0.796)	A	3.048	0.012	0.862	18.08
Macenko	0.787 ± 0.010	(0.774-0.799)	A	2.193	0.018	0.812	17.28
Reinhard	0.773 ± 0.005	(0.764-0.777)	B	0.328	0.309	0.843	15.20
StainNet	0.809 ± 0.009	(0.798-0.820)	A	4.881	0.084	0.951	25.75
StainGAN	0.792 ± 0.008	(0.782-0.802)	A	2.994	0.013	0.794	19.52

anced H&E staining with a clear distinction between purple-blue nuclei and pink cytoplasmic regions. The tissue occupies the frame with minimal background, providing robust material for stain vector estimation. The image shows even illumination with moderate brightness and contrast, free from overexposure, artifacts, or scanning defects.

All classical stain normalization methods evaluated in this study are defined with respect to a single reference image and do not natively support multi-reference normalization. Conversely, the learning-based approaches, such as StainGAN and StainNet, implicitly consider the variability across multiple reference images, rather than by normalization to a single explicit target at inference time. Hence, comparisons involving multiple reference images are considered outside the scope of this evaluation.



(a) Target image **(b)** Unnormalized **(c)** Black and white

(d) Ruifrok **(e)** Vahadane **(f)** Macenko

(g) Reinhard **(h)** StainNet **(i)** StainGAN

Figure 2. Target image and example source image for the different normalization methods

Figure 2 presents the same source image, `TUM-YYWKMWNE.tif`, normalized with the different approaches. For benchmarking purposes, a black-and-white version of the image is included in the analysis, as shown in Figure 2c.

The metrics to evaluate the reduction in color bias with each normalization technique are shown in Table 5. The 95% confidence intervals for classification accuracy were calculated using the Student's t-distribution with $n - 1$ degrees of freedom based on the 5-fold cross-validation accuracy scores, whilst the Critical Letter Difference (CLD) was determined using the Friedman test followed by the Nemenyi post-hoc multiple comparison procedure, where methods sharing the same letter are not significantly different at $\alpha = 0.05$.

The unnormalized dataset achieves high color-only classification accuracy, indicating that a simple classifier trained solely on color statistics can predict tissue classes without accessing biological structure. Color normalization aims to remove exactly this type of bias.

Conversely, black-and-white conversion serves as a theoretical lower bound for color-bias removal because it eliminates all chromatic information. Once every image is reduced to luminance values, no class can be distinguished by color, only by actual tissue structure. However, grayscale images cannot be used in real-world pathology workflows because clinically meaningful stain contrasts are lost. As a reference point, the black-and-white illustration shows how much of the original dataset's separability is due to color alone and provides a conceptual limit against which the effectiveness of practical normalization methods can be evaluated.

All stain normalization methods reduce color accuracy relative to unnormalized images, with Reinhard achieving the lowest color accuracy, suggesting better color bias mitigation.

All Wasserstein and MMD values are computed using the same color-only feature representation and identical kernel settings across normalization methods. Nevertheless, Reinhard presents a high Mean MMD value, indicating suboptimal distribution alignment between normalized images and references, despite showing the lowest Mean Wasserstein distance. This suggests that while Reinhard normalizes some aspects of the color distribution effectively, it may distort others, leading to a higher statistical discrepancy in MMD terms.

While the Wasserstein distance indicates that Reinhard normalization reduces global chromatic shifts across tissue classes, the higher MMD values reveal that local distributional differences persist. This disparity suggests that Reinhard aligns the centroids of the class color distributions but does not harmonize their internal structure. In practice, this means that convolutional models may still exploit fine-grained stain or tissue-specific color signatures after Reinhard normalization. Therefore, the low Wasserstein values should not be interpreted as evidence that color has been completely miti-

gated as a confounding factor.

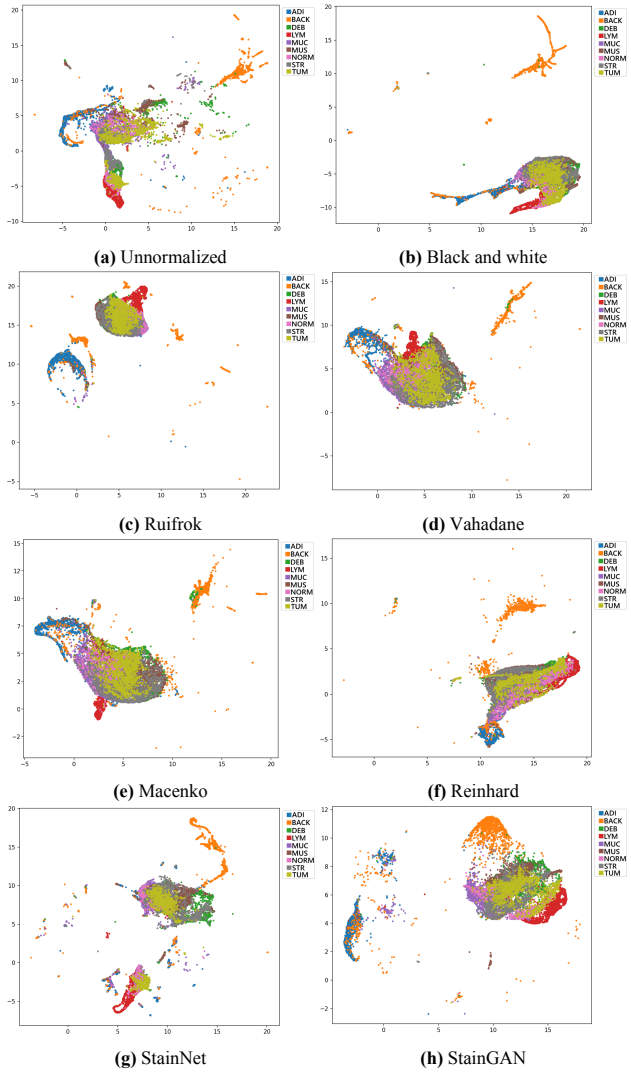


Figure 3. UMAP visualization of feature embeddings colored by tissue class for different normalization methods.

The other classical approaches, namely Macenko, Vahadane, and Ruifrok, reduce inter-class Wasserstein and MMD distances and lower the color-only accuracy, indicating a stain harmonization.

In contrast, deep-learning approaches preserve visual appearance well, achieving the highest SSIM and PSNR values, but do not substantially diminish color-based separability. Thus, the results indicate that normalization reduces color bias, but no method can eliminate it from the NCT-CRC-HE-100K-NONORM dataset.

Because these metrics indicate that color bias persists across methods, we use Uniform Manifold Approximation and Projection (UMAP) projections to explore how this bias distributes across classes and to determine which tissue types remain most easily distinguishable.

The UMAP projection is computed using the RGB channel statistics, perceptually uniform Lab color statistics, and Optical Density (OD) features. These descriptors summarize raw color, perceptual chromaticity, and stain absorbance, and provide a representation of how these color characteristics cluster by class, revealing the persistent color bias.

Figure 3 shows that most tissue classes collapse into an overlapping manifold, especially under the classical methods. Their separation in the unnormalized dataset appears to be due to stain-induced color differences that normalization removes. In contrast, BACK and ADI classes remain outliers regardless of the method applied, demonstrating that their distinctiveness is robust to stain harmonization and reflects intrinsic properties rather than removable color bias. Thus, the observed color signatures primarily arise from these specific classes.

This observation is corroborated by Figure 4, which displays inter-class centroid distances in the unified UMAP embedding space and illustrates the global separation between tissue classes across all normalization methods.

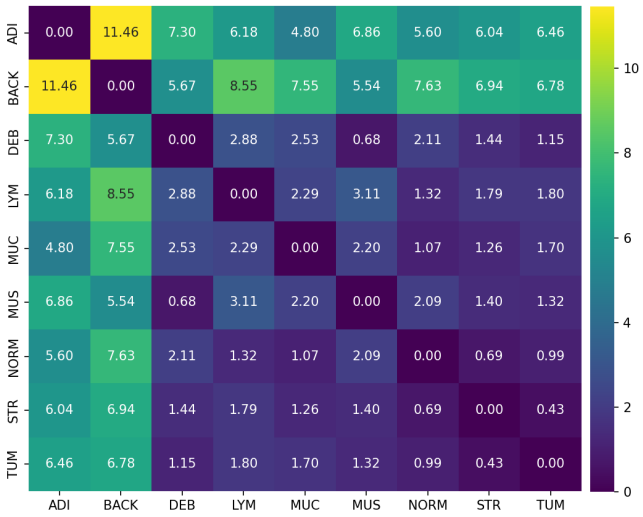


Figure 4. Inter-class centroid distance matrix in unified UMAP space

Larger values in Figure 4 indicate stronger separability between classes. As a consequence, BACK and ADI may be interpreted as the most distinct. Their mutual distance is the highest in the matrix, and both show large distances to all other classes.

In contrast, the other tissue classes exhibit smaller distances between one another, suggesting that once color variation is reduced, many tissue types become difficult to distinguish based solely on the color-derived features used in the UMAP.

The global silhouette score when clustering by normalization method is 0.115, suggesting that the different techniques produce technically distinct outputs. Conversely, the silhouette score for clustering by tissue class is -0.051, indicating that, overall, the normalized color features fail to form distinct clusters by biological class. This is a desirable outcome, as it demonstrates that color-based bias has been reduced across most tissue types. However, Figure 4 indicates that two classes, ADI and BACK, remain separated from the others and from each other.

This demonstrates that normalization mitigates spurious color biases for the majority of classes, but it cannot eliminate the color signature of these specific tissues.

This is because stain variability in histopathology arises from interactions among biological, chemical, and scanner-dependent factors, which cannot be completely homogenized by post-processing normalization. As a result, these classes remain well separated in feature space even after normalization, indicating that a portion of the color signature encodes

Table 6. Classification performance comparison of normalization methods.

Method	Acc.	W. Precision	W. F1
Unnormalized	0.910 ± 0.007	0.910 ± 0.007	0.910 ± 0.007
B/W	0.895 ± 0.007	0.898 ± 0.007	0.896 ± 0.007
Ruifrok	0.900 ± 0.007	0.904 ± 0.007	0.901 ± 0.007
Vahadane	0.895 ± 0.007	0.896 ± 0.007	0.894 ± 0.007
Macenko	0.893 ± 0.007	0.894 ± 0.007	0.892 ± 0.007
Reinhard	0.894 ± 0.007	0.896 ± 0.007	0.894 ± 0.007
StainNet	0.905 ± 0.007	0.907 ± 0.007	0.905 ± 0.007
StainGAN	0.894 ± 0.007	0.894 ± 0.007	0.893 ± 0.007

irreducible, biologically grounded differences rather than correctable technical artifacts.

4.1.2 Impact on classification performance

To quantitatively evaluate the impact of stain normalization on classification performance, we trained and tested an EfficientNetB0 Convolutional Neural Network (CNN) on the different unnormalized and normalized versions of the dataset [Tan and Le, 2019]. The evaluation protocol uses a fixed 70/15/15 train-validation-test split with the same subset of data and the same preprocessing.

The model architecture consists of the base EfficientNetB0 with global average pooling, followed by a dropout layer (0.4 rate) and a final dense softmax classification layer. The model was trained for 12 epochs with an Adam optimizer (learning rate 1e-4), batch size of 32, and incorporated early stopping and learning rate reduction on plateau. The dataset was split into 70% training, 15% validation, and 15% test sets with stratification.

The performance of each method, evaluated by accuracy, weighted precision, and weighted F1 score, is compared in Table 6. The values correspond to the mean performance on the fixed test set, with standard deviations estimated via bootstrap resampling (2,000 iterations). The CNN demonstrated no significant performance differences across the various normalization techniques, as evidenced by two-proportion z-tests on the F1 scores achieved with all methods in comparison to the unnormalized dataset, followed by Holm correction for multiple comparisons. After correction, all pairwise comparisons yielded adjusted p-values equal to 1.0, indicating no statistically significant differences between normalization strategies.

Table 7. F1 scores across different normalization methods

Method	ADI	BACK	Avg other classes
Unnormalized	0.97	0.99	0.89
B/W	0.99	0.99	0.87
Ruifrok	0.98	0.99	0.88
Vahadane	0.97	0.98	0.87
Macenko	0.97	0.98	0.87
Reinhard	0.98	0.99	0.87
StainNet	0.98	0.99	0.88
StainGAN	0.99	0.99	0.87

This finding is consistent with prior studies showing that there is no statistically significant difference in the effectiveness of stain-normalized and non-normalized histopathological images [Voon *et al*., 2023].

A biased behavior, however, is observed in the model, favoring the classes ADI and BACK, as seen in Table 7. This is likely due to the discriminability of these classes seen in Figure 4, which suggests that the model leverages these more distinct visual features. This class-specific bias persists across all normalization methods, suggesting that the performance patterns are caused by dataset characteristics rather than normalization artifacts.

Table 8. Global and class-conditional ECE

Method	Global	ADI	BACK
Unnormalized	0.061	0.022	0.024
B/W	0.058	0.023	0.020
Ruifrok	0.057	0.019	0.031
Vahadane	0.054	0.029	0.028
Macenko	0.064	0.012	0.040
Reinhard	0.057	0.025	0.038
StainNet	0.061	0.026	0.022
StainGAN	0.073	0.032	0.052

This bias is reinforced by Table 8, which presents the Expected Calibration Error (ECE) for each normalization method. Globally, calibration remains relatively stable across methods, indicating that normalization does not substantially affect overall confidence reliability. However, ADI and BACK exhibit lower ECE values compared to the global metric, reflecting more reliable confidence estimates for these classes.

Despite the reductions in color-based separability observed in Section 4.1, these differences did not produce meaningful changes in the CNN performance. This is expected, as the metrics in Table 5 quantify color bias in isolation using color-only descriptors, whereas the EfficientNetB0 model learns a richer set of morphological and textural features that are not affected by normalization.

4.1.3 Blue dominance in stain normalization

The DEB class of the NCT-CRC-HE-100K dataset suffers from hematoxylin oversaturation, producing a strong blue-purple cast and suppressing eosin contrast, likely resultant from the color normalization process [Ignatov and Malivenko, 2024].

The Macenko algorithm, used in the original dataset, may introduce blue artifacts in the source image [Khan *et al*., 2025].

The propagation of these color artifacts may happen, for instance, if the chosen reference is suboptimal [Jawad and Khursheed, 2024].

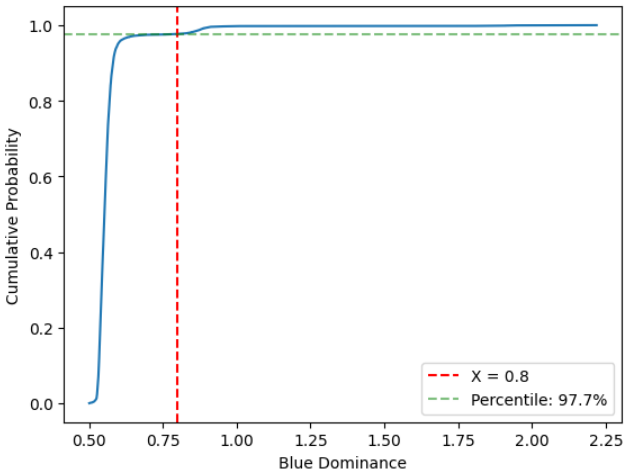
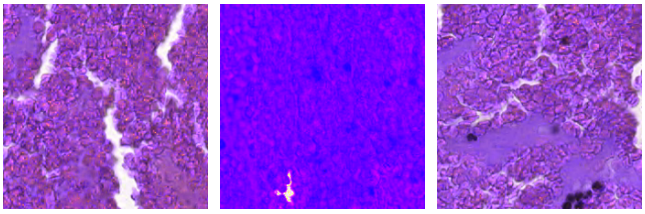
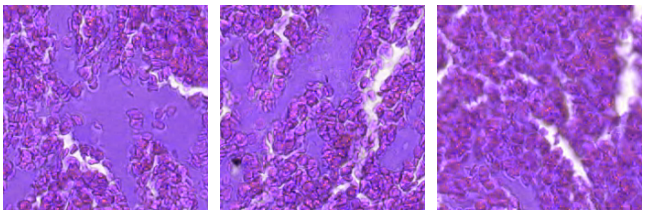


Figure 5. Empirical ECDF of Blue Dominance scores

The empirical Cumulative Distribution Function (CDF) plot, depicted in Figure 5, shows that a BD value of 0.8 lies in the upper tail of the distribution of the class DEB in the original Macenko dataset. At this threshold, the cumulative probability is approximately 97.7%, meaning that about 2.3% of all samples exceed it. Thus, 0.8 is used as a threshold to detect highly blue-dominant images in the datasets.



(a) DEB-GFNEVMQL.tif **(b)** DEB-THAKCKKD.tif **(c)** DEB-CCEHQYAV.tif



(d) DEB-AILNQVEC.tif **(e)** DEB-APPITQWG.tif **(f)** DEB-CARVSHCH.tif

Figure 6. Images in DEB class with blue color over saturation in the original dataset

Examples of stain-normalized images with extreme blue dominance values are shown in Figure 6. Figure 6a illustrates an example at the blue dominance threshold of 0.80. Conversely, Figure 6b represents the highest BD score observed in the dataset (2.22). The image in Figure 2f, as a reference, has a BD score of 0.59.

Saturation artifacts observed in the DEB class have been associated with blue-channel clipping caused by intensity overflow during color normalization, which manifests as pixels capped at the maximum 8-bit value (255) Ignatov and Malivenko [2024]. However, this criterion alone lacks sensitivity to detect the range of color distortions introduced by normalization. As examples, approximately 99.85% of the pixels in the image shown in Figure 6b have the blue channel capped at the maximum value of 255, whereas only 8.35% of the pixels in Figure 6a reach this limit, despite this image lying in the upper tail of the blue-dominance distribution shown in Figure 5.

It is also noted that for the original Macenko implementation and for Ruifrok, the DEB class exhibits a higher sensitivity to blue over-representation than the dataset as a whole, indicating that color artifacts introduced by these normalization methods are not uniformly distributed across tissue types. However, blue-dominant images are not exclusive to this class and are also observed in other tissue classes, although in lower frequencies. Conversely, other methods are able to suppress extreme blue dominance for both DEB and the full dataset.

Table 9. Blue dominant images across color normalization methods

Method	**DEB (BD ≥ 0.8)**	**All classes (BD ≥ 0.8)**
Unnormalized	0%	0%
Macenko (original)	2.30%	0.28%
Ruifrok	14.00%	8.52%
Vahadane	0%	2.21%
Macenko	0%	0.02%
Reinhard	0%	0%
StainNet	0%	0.75%
StainGAN	0%	0%

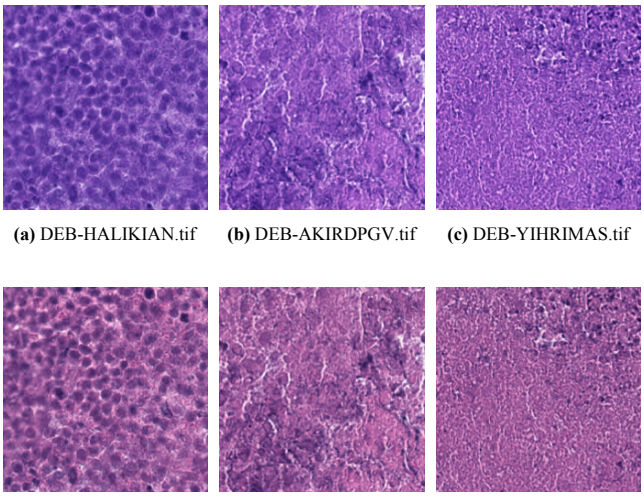
As noted from Table 9, the DEB class of the original NCT-CRC-HE-100K dataset contains 2.30% of the images with greater blue dominance than that of Figure 6a. No blue-dominant images are detected when the Macenko method is re-applied in this study from NCT-CRC-HE-100k-NONORM. This indicates that the observed saturation is not a limitation of the Macenko algorithm itself, but a consequence of the specific normalization configuration or reference selection used during the dataset's original preprocessing.

Furthermore, among the tested color normalization techniques in this work, Ruifrok returns the highest number of images with a blue dominance score exceeding the threshold. This can also be observed in Figure 2d, which appears bluer than the resulting images from the other normalization methods.

This blue saturation in Ruifrok is shown in Figure 7 alongside the same images from the Macenko normalization. For `DEB-HALIKIAN.tif`, a blue dominance of 0.64 is achieved in Macenko and 0.90 for Ruifrok; for `DEB-AKIRDPGV.tif`, Macenko achieves a 0.62 score, whereas Ruifrok results in 0.81. Ultimately, `DEB-YIHRIMAS.tif` has a blue dominance score of 0.63 for Macenko and 0.81 for Ruifrok.

A greater proportion of blue-dominant images is found in Ruifrok than in the original dataset. This, however, occurs more homogeneously across Ruifrok images, reducing potential biases in subsequent analysis.

The statistical evidence presented in Table 10 corroborates this observation. While Ruifrok normalization produces a higher mean blue dominance score than the unnormalized images and the original Macenko dataset, it maintains a lower



(a) DEB-HALIKIAN.tif **(b)** DEB-AKIRDPGV.tif **(c)** DEB-YIHRIMAS.tif

(d) DEB-HALIKIAN.tif **(e)** DEB-AKIRDPGV.tif **(f)** DEB-YIHRIMAS.tif

Figure 7. Ruifrok (top row) blue dominance in comparison with Macenko (bottom row)

standard deviation and a shorter min-max range than the original dataset. This suggests that Ruifrok introduces a more homogeneous color shift, rather than unpredictable artifacts, thus reducing unpredictability compared to the original dataset configuration.

Table 10. Blue dominance score statistics per normalization method (all classes aggregated)

Normalization method	Min	Max	Mean	Std
Unnormalized	0.24	0.68	0.52	0.05
Macenko (original)	0.24	2.22	0.56	0.08
Ruifrok	0.50	1.68	0.71	0.17
Vahadane	0.51	0.88	0.60	0.06
Macenko	0.51	0.83	0.59	0.03
Reinhard	0.58	0.60	0.59	0.01
StainNet	0.43	0.89	0.57	0.07
StainGAN	0.48	0.72	0.58	0.04

This consistent normalization effect is visually confirmed in Figure 3. The UMAP projection reveals that Ruifrok successfully aggregates all other classes, including DEB, into a coherent cluster, effectively separating them from ADI and BACK. This level of class separation is comparable to that achieved by other normalization methods.

Conversely, the unpredictable color distribution in the original dataset is a challenge for computational pathology. The unpredictability observed in the DEB class may impede model generalization, as algorithms may learn spurious color-based correlations rather than robust morphological features.

4.2 Distortion-based bias analysis

To evaluate the presence of distortion-related artifacts in the CRC-HE-100K dataset, we compute the BRISQUE score for all images across the nine tissue classes.

Figure 8 shows the distribution of BRISQUE values per tissue class, with a noticeable heterogeneity in image quality across classes. The BACK class exhibits the highest median BRISQUE scores, suggesting that this group contains a larger proportion of images affected by distortions. Examples of these lower-quality images in BACK are shown in Figure 9.

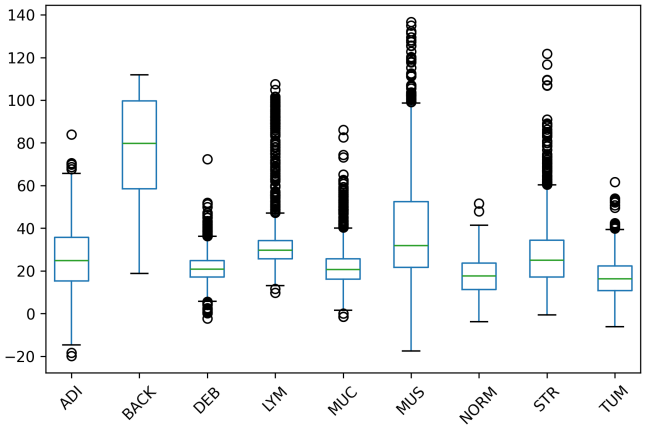
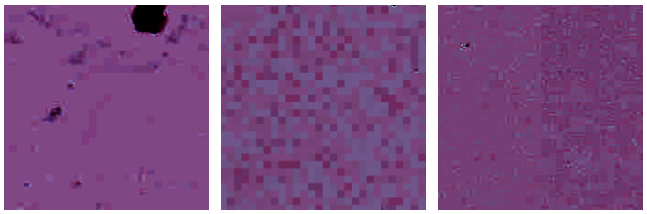


Figure 8. Distribution of BRISQUE scores across tissue classes



(a) BACK-AECPLTMV.tif **(b)** BACK-FLIHYDKI.tif **(c)** BACK-PDVTCLTQ.tif

Figure 9. Examples of images in BACK class with distortion artifacts

In contrast, the other classes exhibit lower BRISQUE medians, indicating fewer detectable distortion anomalies. Therefore, the analysis demonstrates that these artifacts are not uniformly distributed across the dataset. This non-uniformity introduces a potential source of technical bias, as models may learn class-specific noise patterns rather than biologically meaningful morphology [Dehkharghanian *et al.*, 2023].

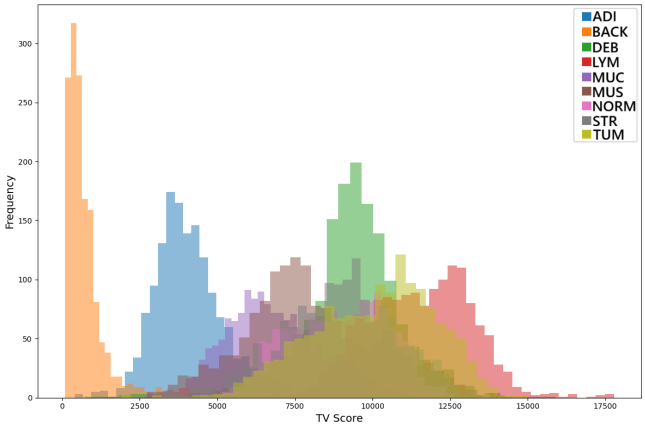


Figure 10. Total Variation distribution per class

A Kruskal–Wallis test confirmed that the nine tissue classes do not share a common BRISQUE distribution ($H = 5749.13$, $p < 0.001$), indicating firm global heterogeneity in distortion-related artifacts. Post hoc Dunn tests with Bonferroni correction showed that almost all tissue classes differed significantly from one another. Only three class pairs did not exhibit statistically significant differences: ADI–STR ($p = 0.75$), DEB–MUC ($p = 1.00$), and NORM–TUM ($p = 0.45$), suggesting that these groups share similar levels of artifacts. In contrast, the BACK class differed from all others with p-values as low as 10^{-165}, confirming that BACK contains significantly higher BRISQUE scores. These results reinforce the finding that distortion artifacts are highly class-dependent in

this dataset, with BACK being the most severely affected.

The greater BRISQUE score for BACK is consistent with the observation that this class includes large homogeneous regions where processing artifacts become more perceptible. As shown in Figure 10, the BACK class has significantly lower TV values than other tissue classes.

Correlation analyses between BRISQUE and TV scores evidence the statistical significance of these quality disparities. There exists a moderate negative Pearson correlation ($r = -0.495$, $p < 0.001$) and Spearman correlation ($\rho = -0.351$, $p < 0.001$), quantitatively confirming that images with lower texture complexity in the dataset tend to exhibit more pronounced distortion artifacts.

Hence, it is noted that the BACK class contains significantly more processing artifacts than the other classes of the original dataset, thereby introducing a potential confounding signal.

4.2.1 Isolation of high-frequency distortion cues

To investigate whether compression-related distortion artifacts contribute to the observed class-specific bias, we isolate distortion-related cues using FFT-based high-pass filtering to retain high-frequency content, while low-pass filtering was applied to suppress these components and preserve lower-frequency morphological structure. This is because compression artifacts are characterized by abrupt local intensity transitions, which predominantly contribute to high-frequency components in the spatial frequency spectrum.

For each normalization method, BRISQUE-based distortion descriptors were extracted from high-pass and low-pass filtered images, and a classifier was trained to assess whether high-frequency distortion cues alone are sufficient to sustain classification performance.

Table 11 reports the resulting accuracies. It is important to note that these accuracies are lower than those obtained in Section 4.1.2, as the classifier in this experiment is trained exclusively on distortion-related descriptors, and not on the full image content.

Across all normalization methods, high-pass filtering reduced performance relative to the original condition, which indicates that high-frequency components alone are insufficient to maintain classification accuracy. In contrast, low-pass filtering preserved performance to a large extent. Since low-pass filtering attenuates high-frequency components associated with compression artifacts, these results suggest that distortion-related signals are not the primary features exploited by the classifier.

Taken together, these findings indicate that although the BACK class exhibits elevated distortion metrics, the observed class-specific bias cannot be attributed solely to high-frequency artifact content. Instead, the persistence of classification performance under low-pass filtering supports the interpretation that intrinsic morphological and structural characteristics of the dataset are the dominant drivers of the bias, rather than compression-induced distortions.

Thus, while it remains possible that high-capacity models may partially leverage subtle artifact-related cues, our controlled frequency-isolation experiment indicates that such cues are not sufficient to account for the class-specific bias observed in this dataset. Furthermore, no normalization method sustained performance under high-pass filtering. This indicates that the elevated distortion metrics observed for the BACK class do not imply that discriminability is attributable to distortion.

Table 11. Classification accuracy under frequency isolation conditions

Normalization method	Filter type	Accuracy
B/W	Original	0.563
	High-pass	0.371
	Low-pass	0.521
Macenko (original)	Original	0.554
	High-pass	0.400
	Low-pass	0.567
Ruifrok	Original	0.492
	High-pass	0.379
	Low-pass	0.475
Vahadane	Original	0.513
	High-pass	0.354
	Low-pass	0.525
Macenko	Original	0.396
	High-pass	0.346
	Low-pass	0.504
Reinhard	Original	0.504
	High-pass	0.392
	Low-pass	0.542
StainNet	Original	0.571
	High-pass	0.433
	Low-pass	0.525
StainGAN	Original	0.546
	High-pass	0.333
	Low-pass	0.563

5 Conclusions and future works

This work presents a quantitative assessment of color bias, stain-normalization behavior, and distortion-related artifacts in the widely used NCT-CRC-HE-100K and NCT-CRC-HE-100K-NONORM histopathology datasets. The analysis extended previous observations by providing reproducible evidence that low-level image characteristics can significantly influence tissue class separability in this dataset.

It demonstrates that all evaluated normalization methods reduced, but did not eliminate, class-dependent color signatures. The UMAP projections further revealed persistent clustering patterns, with BACK and ADI consistently forming distinct manifolds irrespective of normalization approach.

Complementing these findings, the CNN classification experiment showed that model performance is maintained stable across all normalization techniques, including the unnormalized version. This indicates that even when color separability decreases, convolutional networks still extract discriminative information from morphological patterns.

Additionally, the analysis on the DEB class confirmed the presence of hematoxylin oversaturation and its amplification under specific normalization strategies. While the original dataset exhibited heterogeneous blue-dominance artifacts, Ruifrok normalization introduced a more uniform chromatic

shift.

Ultimately, the assessment of distortion artifacts showed a non-uniform distribution across tissue classes, with BACK containing significantly more low-quality regions, possibly due to image compression. The combination of high BRISQUE scores and low Total Variation suggests that these smooth background patches are particularly vulnerable to distortion effects. Although distortion artifacts are unevenly distributed across classes, frequency isolation experiments indicate that high-frequency artifact cues alone are insufficient to explain class-specific classification bias.

These results reinforce the need for careful dataset curation and more robust benchmarking practices in computational pathology. The persistence of color and distortion biases implies that high classification accuracies reported on NCT-CRC-HE-100K may overestimate the biological validity and generalizability of trained models.

Future works may examine whether the observed color inconsistencies, distortion artifacts, and class-dependent biases generalize to other datasets, such as LC25000 or CRC-ICM. Further research could also investigate whether previously reported high-accuracy models maintain their performance as dataset biases are mitigated. It is also proposed to use a multi-reference or adaptive target color normalization on a similar experiment.

Acknowledgements

The authors would like to thank the support of the University of Brasilia.

Funding

This research received no funding

Authors' Contributions

G.A.P.R.: Conceptualization, Formal Analysis, Investigation, Writing – Original Draft. A.L.M.S.: Validation, Visualization. G.P.R.F., R.B., V.P.G. and M.R.: Methodology, Writing – Review & Editing. R.I.M.: Conceptualization, Supervision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset used in this study is publicly available at `https://zenodo.org/records/1214456` [Kather *et al*., 2018].

References

Barbano, C. A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., and Grangetto, M. (2021). Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 76–80. IEEE. DOI: 10.1109/icip42928.2021.9506198.

Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., and Mastorides, S. M. (2019). Lung and colon cancer histopathological image dataset (LC25000). *arXiv preprint arXiv:1912.12142*. DOI: 10.48550/arXiv.1912.12142.

Cordova, R., Viallon, V., Fontvieille, E., Peruchet-Noray, L., Jansana, A., Wagner, K.-H., Kyrø, C., Tjønneland, A., Katzke, V., Bajracharya, R., *et al*. (2023). Consumption of ultra-processed foods and risk of multimorbidity of cancer and cardiometabolic diseases: a multinational cohort study. *The Lancet Regional Health–Europe*, 35. DOI: 10.1016/j.lanepe.2023.100771.

Dehkharghanian, T., Bidgoli, A. A., Riasatian, A., Mazaheri, P., Campbell, C. J., Pantanowitz, L., Tizhoosh, H., and Rahnamayan, S. (2023). Biased data, biased AI: deep networks predict the acquisition site of tcga images. *Diagnostic pathology*, 18(1):67. DOI: 10.1186/s13000-023-01355-3.

Di Giammarco, M., Martinelli, F., Santone, A., Cesarelli, M., and Mercaldo, F. (2024). Colon cancer diagnosis by means of explainable deep learning. *Scientific reports*, 14(1):15334. DOI: 10.1038/s41598-024-63659-8.

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., and Kompatsiaris, I. (2022). A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552. DOI: 10.1016/j.cviu.2022.103552.

Filho, A. M., Laversanne, M., Ferlay, J., Colombet, M., Piñeros, M., Znaor, A., Parkin, D. M., Soerjomataram, I., and Bray, F. (2025). The globocan 2022 cancer estimates: data sources, methods, and a snapshot of the cancer burden worldwide. *International Journal of Cancer*, 156(7):1336–1346. DOI: 10.1002/ijc.35278.

Firildak, K., Celik, G., and Talu, M. F. (2025). Supervised constructive learning-based model for identifying colorectal cancer tissue types from histopathological images. *International Journal of Imaging Systems and Technology*, 35(4):e70161. DOI: 10.1002/ima.70161.

Hoque, M. Z., Keskinarkaus, A., Nyberg, P., and Seppänen, T. (2024). Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102:101997. DOI: 10.1016/j.inffus.2023.101997.

Ignatov, A. and Malivenko, G. (2024). NCT-CRC-HE: Not all histopathological datasets are equally useful. In *European Conference on Computer Vision*, pages 300–317. Springer. DOI: 10.48550/arXiv.2409.11546.

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., and Madabhushi, A. (2019). Histoqc: an open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics*, 3:1–7. DOI: 10.1200/cci.18.00157.

Jawad, M. A. and Khursheed, F. (2024). A novel approach for color-balanced reference image selection for breast histology image normalization. *Biomedical Signal Processing and Control*, 94:106299. DOI: 10.21203/rs.3.rs-3833711/v1.

Jiang, X., Hu, Z., Wang, S., and Zhang, Y. (2023). Deep learning for medical image-based cancer diagnosis. *Cancers*, 15(14):3608. DOI: 10.3390/cancers15143608.

Juul, F., Parekh, N., Martinez-Steele, E., Monteiro, C. A., and Chang, V. W. (2022). Ultra-processed food consumption among us adults from 2001 to 2018. *The American journal of clinical nutrition*, 115(1):211–221. DOI: 10.1093/ajcn/nqab305.

Kang, H., Luo, D., Feng, W., Zeng, S., Quan, T., Hu, J., and Liu, X. (2021). Stainnet: a fast and robust stain normalization network. *Frontiers in Medicine*, 8:746307. DOI: 10.3389/fmed.2021.746307.

Kastryulin, S., Zakirov, J., Prokopenko, D., and Dylov, D. V. (2022). Pytorch image quality: Metrics for image quality assessment. *arXiv preprint arXiv:2208.14818*. DOI: 10.2139/ssrn.4206741.

Kather, J. N., Halama, N., and Marx, A. (2018). 100,000 histological images of human colorectal cancer and healthy tissue (v0.1). DOI: 10.5281/zenodo.1214456.

Khan, U., Härkönen, J., Friman, M., Latonen, L., Kuopio, T., and Ruusuvuori, P. (2025). Staining normalization in histopathology: Method benchmarking using multicenter dataset. *arXiv preprint arXiv:2506.19106*. DOI: 10.1038/s41598-026-40943-3.

Kheiri, F., Rahnamayan, S., Makrehchi, M., and Asilian Bidgoli, A. (2025). Investigation on potential bias factors in histopathology datasets. *Scientific Reports*, 15(1):11349. DOI: 10.1038/s41598-025-89210-x.

Levy, R. B., Barata, M. F., Leite, M. A., and Andrade, G. C. (2024). How and why ultra-processed foods harm human health. *Proceedings of the Nutrition Society*, 83(1):1–8. DOI: 10.1017/s0029665123003567.

Li, M. (2024). Transformer-based self-supervised learning and distillation for medical image classification: Improving colorectal cancer detection on nct-crc-he-100k with swin-t v2. In *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, pages 644–648. IEEE. DOI: 10.1109/cbase64041.2024.10824558.

Liu, Z. and He, K. (2024). A decade's battle on dataset bias: Are we there yet? *arXiv preprint arXiv:2403.08632*. DOI: 10.48550/arxiv.2403.08632.

Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC. DOI: 10.2307/1271491.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and 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*, pages 1107–1110. IEEE. DOI: 10.1109/isbi.2009.5193250.

Meine, G. C., Picon, R. V., Santo, P. A. E., and Sander, G. B. (2024). Ultra-processed food consumption and gastrointestinal cancer risk: A systematic review and meta-analysis. *Official journal of the American College of Gastroenterology| ACG*, 119(6):1056–1065. DOI: 10.14309/ajg.0000000000002826.

Merabet, A., Saighi, A., Saad, H., Ferradji, M. A., Laboudi, Z., Almaktoom, A. T., Mousavirad, S. J., Elbatal, I., and Mohamed, A. W. (2025). Ai for colon cancer: a focus on classification, detection, and predictive modeling. *International Journal of Medical Informatics*, page 106115. DOI: 10.1016/j.ijmedinf.2025.106115.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708. DOI: 10.1109/tip.2012.2214050.

Mokhtari, Z., Amjadi, E., Bolhasani, H., Faghih, Z., Dehghanian, A., and Rezaei, M. (2023). Crc-icm: Colorectal cancer immune cell markers pattern dataset. *arXiv preprint arXiv:2308.10033*. DOI: 10.48550/arxiv.2308.10033.

Pacal, I. and Attallah, O. (2025). Hybrid deep learning model for automated colorectal cancer detection using local and global feature extraction. *Knowledge-Based Systems*, page 113625. DOI: 10.1016/j.knosys.2025.113625.

Pocock, J., Graham, S., Vu, Q. D., Jahanifar, M., Deshpande, S., Hadjigeorghiou, G., Shephard, A., Bashir, R. M. S., Bilal, M., Lu, W., *et al*. (2022). Tiatoolbox as an end-to-end library for advanced tissue image analytics. *Communications medicine*, 2(1):120. DOI: 10.1038/s43856-022-00186-5.

Popkin, B. M. and Laar, A. (2025). Nutrition transition's latest stage: Are ultra-processed food increases in low-and middle-income countries dooming our preschoolers' diets and future health? *Pediatric Obesity*, 20(5):e70002. DOI: 10.2139/ssrn.4872344.

Qin, Z., Sun, W., Guo, T., and Lu, G. (2024). Colorectal cancer image recognition algorithm based on improved transformer. *Discover Applied Sciences*, 6(8):422. DOI: 10.1007/s42452-024-06127-2.

Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2002). Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41. DOI: 10.1109/38.946629.

Rezaei, M., Amjadi, E., Bolhasani, H., Dehghanian, A., Sanei, M., Faghih, Z., *et al*. (2023). Colorectal cancer immune cell markers dataset v1 (crc-icm-v1). Avaialble at:`https://data.mendeley.com/datasets/h3fhg9zr47/2`.

Rinaldi, A. M., Russo, C., and Tommasino, C. (2022). Effects of color stain normalization in histopathology image retrieval using deep learning. In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 26–33. IEEE. DOI: 10.1109/ism55400.2022.00010.

Rodrigues, R. M., Souza, A. d. M., Bezerra, I. N., Pereira, R. A., Yokoo, E. M., and Sichieri, R. (2021). Evolução dos alimentos mais consumidos no brasil entre 2008-2009 e 2017-2018. *Revista de Saúde Pública*, 55:4s.

Roy, S., kumar Jain, A., Lal, S., and Kini, J. (2018). A study about color normalization methods for histopathology images. *Micron*, 114:42–61. DOI: 10.1016/j.micron.2018.07.005.

Ruifrok, A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*. Available at:`https://pubmed.ncbi.nlm.nih.gov/11531144/`.

Shaban, M. T., Baur, C., Navab, N., and Albarqouni, S. (2019). Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE. DOI: 10.1109/isbi.2019.8759152.

Shi, L., Li, X., Hu, W., Chen, H., Chen, J., Fan, Z., Gao, M., Jing, Y., Lu, G., Ma, D., *et al*. (2023). EBHI-

Seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10:1114673. DOI: /10.3389/fmed.2023.1114673.

Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., *et al*. (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502. DOI: 10.1016/j.media.2016.08.008.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. DOI: 10.48550/arxiv.1905.11946.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and Van Der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544. DOI: 10.1016/j.media.2019.101544.

Uddin, A. H., Chen, Y.-L., Akter, M. R., Ku, C. S., Yang, J., and Por, L. Y. (2024). Colon and lung cancer classification from multi-modal images using resilient and efficient neural network architectures. *Heliyon*, 10(9). DOI: 10.1016/j.heliyon.2024.e30625.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971. DOI: 10.1109/tmi.2016.2529665.

Voon, W., Hum, Y. C., Tee, Y. K., Yap, W.-S., Nisar, H., Mokayed, H., Gupta, N., and Lai, K. W. (2023). Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images. *Scientific Reports*, 13(1):20518. DOI: 10.1038/s41598-023-46619-6.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O. (2022a). Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810. DOI: 10.1007/s11263-022-01625-5.

Wang, L., Du, M., Wang, K., Khandpur, N., Rossato, S. L., Drouin-Chartier, J.-P., Steele, E. M., Giovannucci, E., Song, M., and Zhang, F. F. (2022b). Association of ultra-processed food consumption with colorectal cancer risk among men and women: results from three prospective us cohort studies. *bmj*, 378. DOI: 10.1136/bmj-2021-068921.

Xi, Y. and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational oncology*, 14(10):101174. DOI: 10.1016/j.tranon.2021.101174.

Yu, J., Feng, Q., Kim, J. H., and Zhu, Y. (2022). Combined effect of healthy lifestyle factors and risks of colorectal adenoma, colorectal cancer, and colorectal cancer mortality: systematic review and meta-analysis. *Frontiers in oncology*, 12:827019. DOI: 10.3389/fonc.2022.827019.