scientific reports

Check for updates

OPEN

Investigation on potential bias factors in histopathology datasets

Farnaz Kheiri[1]✉, Shahryar Rahnamayan[2], Masoud Makrehchi[1] & Azam Asilian Bidgoli[3]

Deep neural networks (DNNs) have demonstrated remarkable capabilities in medical applications, including digital pathology, where they excel at analyzing complex patterns in medical images to assist in accurate disease diagnosis and prognosis. However, concerns have arisen about potential biases in The Cancer Genome Atlas (TCGA) dataset, a comprehensive repository of digitized histopathology data and serves as both a training and validation source for deep learning models, suggesting that over-optimistic results of model performance may be due to reliance on biased features rather than histological characteristics. Surprisingly, recent studies have confirmed the existence of site-specific bias in the embedded features extracted for cancer-type discrimination, leading to high accuracy in acquisition site classification. This biased behavior motivated us to conduct an in-depth analysis to investigate potential causes behind this unexpected biased ability toward site-specific pattern recognition. The analysis was conducted on two cutting-edge DNN models: KimiaNet, a state-of-the-art DNN trained on TCGA images, and the self-trained EfficientNet. In this research study, the balanced accuracy metric is used to evaluate the performance of a model trained to classify data centers, which was originally designed to learn cancerous patterns, with the aim of investigating the potential factors contributing to the higher balanced accuracy in data center detection.

Keywords Digital pathology, Site-specific bias, Learning models, Histopathology

In the rapidly developing landscape of Artificial Intelligence (AI) and Machine Learning (ML), we are becoming increasingly conscious of an emerging concern - the imperative effects of bias in these systems. As AI technologies revolutionize our lives and play a crucial role in making critical decisions, such as healthcare, recognizing the roots of biased behavior in AI models becomes essential for increasing the trustworthiness and generality of these models[1–4]. Emerging bias in deep models's decision-making process refers to the phenomenon where an AI model is able to attain high accuracy on a particular task without actually learning the features required to deal with that task. In other words, the model relies on heuristics specific to the training data but does not generalize well to unseen data. This can be a significant challenge in real-world machine learning applications[5–7].

In the realm of histopathology, distinct categories of biases that arise from various sources[8,9], as summarized in Fig. 1, can negatively affect definitive diagnosis and prognosis steps. Though specific terminologies have been provided to introduce them, these types of biases are closely related to the concept of 'shortcuts' in the broader machine learning literature. These biases can lead to satisfactory in-distribution performance but poor out-of-distribution performance. The first category, "Sampling bias," occurs when the target population is not randomly picked for the data obtained during model training. In other words, when training data applied to deep models are not impartially representative of the population intended to serve, sampling bias of data kind may appear. Then, the model may not perform equally well for all demographic groups due to underrepresented or overrepresented behavior of data in the training set[10–12]. Roach et al., for the first time, focused on AI prognostic models for digital pathology images from prostate cancer trials and assessed their equity in the context of racial disparities between African American (AA) and non-AA populations, with the dominant portion of the data relating to the non-AA population. Their findings suggest the need for further diverse evaluation in African American cohorts, which is currently ongoing[13]. According to research by DeGrave et al.[14], trained models on radiographic images are more likely to pick up medically unrelated features, which may be caused by bias in data collection than the actual underlying pathology-related information. Another recent research[15] showed that trained models for classifying whole-slide images exhibit significant performance variations among various demographic groups. Dhont et al.[16] assessed five CNNs for automatic COVID-19 screening from chest radiography (CXR). The final performance was impacted by dataset bias, and CNNs were more likely to learn dataset-specific characteristics than patterns relevant to diseases, which calls into question their validity as screening methods on their own. Another research[17] investigated biases in 19 COVID-19 chest X-ray image datasets, highlighting significant

[1]Department of Electrical, Computer and Software Engineering, Ontario Tech University, Oshawa, Canada. [2]Department of Engineering, Brock University, St. Catharines, Canada. [3]Department of Computer Science, Wilfrid Laurier University, Waterloo, Canada. ✉email: farnaz.kheiri@ontariotechu.net

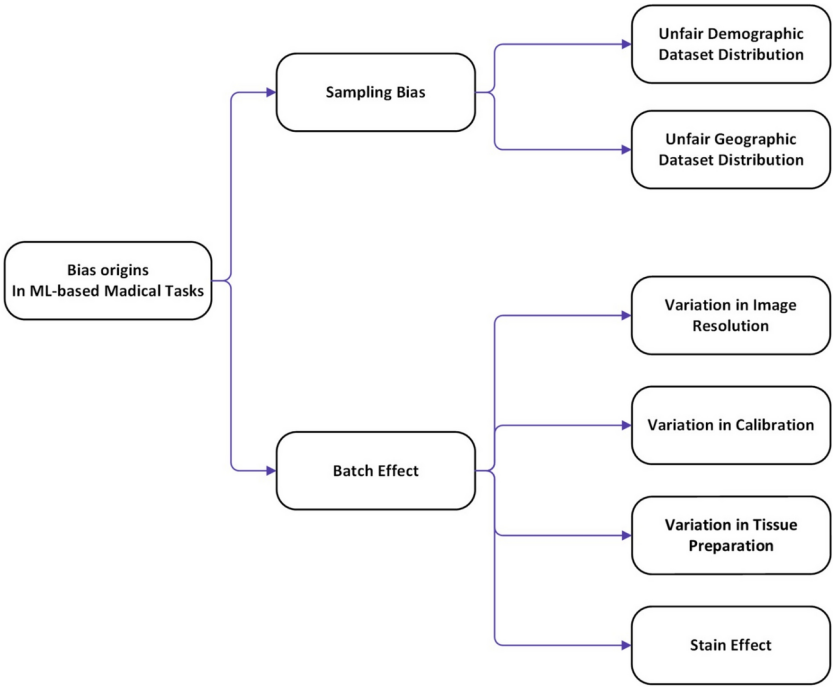


Fig. 1. Potential sources of bias in ML-based medical tasks.

ethical issues despite high classification accuracy in a popular dataset. It emphasizes the need for ethical tools with distribution and image quality considerations to minimize dataset biases and improve future developments in AI-driven COVID-19 diagnostics. Numerous pathology studies have also shown significant health disparities based on other demographic attributes of data, such as gender and age[18]. These disparities can be apparent in several ways, including differences in disease prevalence, treatment outcomes, access to healthcare services, and overall health outcomes among different demographic groups. The discrepancies found in the data may be reproduced or reinforced by AI systems that were trained on biased data or improperly constructed to take demographic variables into account. For instance, if an AI model is biased towards particular ethnic or gender groups as a result of imbalanced data representation or label distribution, it may result in different population areas receiving differing healthcare outcomes[19–21].

The second category, "batch effect," refers to a variation or discrepancy in the measurements of histopathology samples across different data centers. This variation is not related to the specific characteristics or features being studied, but rather, it is affected by environmental factors and the data acquisition process[8,9]. For example, instrumentation variation in data centers, including disparities in hardware specifications, calibration, or image resolution, leads to site-specific differences in image quality and features. Furthermore, variations in data acquisition protocols across sites can introduce differences in staining techniques, tissue preparation, or image preprocessing steps, which may affect the characteristics of the histopathology samples[22,23]. According to Howard et al.[8], the clinical statistics in the TCGA data, such as survival and gene expression patterns, are distributed in startlingly different ways among samples from various clinics and laboratories. They demonstrated that some models detect source sites as opposed to making prognoses or forecasting mutation states. The research paper[24] deals with the problem of skin cancer diagnosis using a technique to identify and measure shortcuts. The work aims to reduce shortcuts and increase the model's reliability for clinical usage in skin cancer diagnosis by reducing bias using image inpainting and retraining the classifier.

A recent study by Dehkharghanian et al.[22] revealed the unexpected capability of TCGA's cancer-type features, extracted by KimiaNet, in predicting 24 acquisition sites. Each acquisition site (data center) contributed more than 1% of the total data samples, providing enough data for meaningful analysis. Using these features, the model achieved nearly 70% accuracy in predicting the acquisition sites. These discoveries demonstrate the presence of embedded site-specific signatures, contributing to the corresponding data centers where WSIs were submitted. Thus, it raises doubts about being biased toward these unrelated signatures for cancer-type detection, which could potentially result in low external validation when dealing with the data collected from unseen data centers. Notably, several studies[25–27] have endeavored slide-based investigations to eliminate these signatures, with the goal of reducing the prediction accuracy of acquisition sites; by doing so, they aim to achieve higher external validation accuracy.

However, before proceeding with any further steps, identifying the origin of these signatures as biases not only prevents the occurrence of suddenly biased results in similar histopathology research but also enhances generalization and trustworthiness toward applying AI models for aiding in diagnosis procedures. Motivated by the conducted experiments aimed at addressing bias issues embedded in the TCGA dataset, we conduct an in-depth analysis of the high-level patch-based TCGA features to systematically show the existence of bias and

investigate potential signatures associated with data centers in cancerous features extracted by deep models. Our research study involves experiments conducted on features from two deep models. The first model is the pretrained KimiaNet, which was trained on the entire TCGA dataset, encompassing 29 cancer types, as described in the "A Summary of DataSet" section. This part of the study focuses on investigating the features extracted by KimiaNet and its associated dataset. In the second part of the experiments, we used EfficientNet. For this part, we narrowed the dataset down to two cancer subtypes, LUAD and LUSC, to fine-tune the model ourselves and test the hypotheses generated from the KimiaNet experiments. The rationale for selecting KimiaNet and EfficientNet stems from previous studies[22,25,26] that investigated bias in TCGA data samples embedded in the features extracted by these models. These studies have explored the challenges of addressing bias in features extracted by KimiaNet and EfficientNet, concluding that fully eliminating bias from a deep model's decision-making process is not straightforward. Additionally, both models have demonstrated high efficiency in pathology-based cancer prediction research, making them suitable for our investigation. The goal of our study is to explore bias factors in the features extracted by these models further using designed experiments, which are categorized into four case studies. However, it's important to note that the tests we apply to the features are model-independent and can be used with features extracted by any learning model.

- **Case Study 1** explores how the data sampling process affects model performance, revealing a distribution dependency between cancer types and data centers in the TCGA dataset. The analysis shows that models may capture site-specific patterns rather than cancer-specific features, highlighting the need to address data distribution biases.
- **Case Study 2** examines the impact of the patching process on bias in the classification of cancer types and data source institutions. The study reveals that co-slide patches significantly influence classification balanced accuracy, leading to slide-specific biases.
- **Case Study 3** investigates site-specific patterns in cancer-based features, showing that patches from the same data center and cancer type share more similarities than those from different centers. This suggests that certain data centers have distinct acquisition protocols, leading to biases in model predictions.
- **Case Study 4** explores the impact of color variations from site-specific staining on model performance by introducing random noise into RGB channels and converting the images to grayscale.

Parts of the bias detection experiments were conducted on the cancerous features already extracted by KimiaNet[28], while the remaining experiments focused on factors that may affect cancerous features during training or pretraining steps. To investigate this further, we trained the EfficientNet model from scratch, aiming to uncover any underlying biases introduced during this stage.

Results

A summary of dataset

In this study, we utilized the TCGA dataset[29], which serves as a comprehensive repository containing 32,072 WSIs encompassing molecular and clinical data originating from 156 data centers with 33 cancer categories. Each WSI, at 20x magnification, comprises an average of 50 tissue patches, each sized 1000×1000 pixels[28]. To ensure a robust dataset with a high fraction of tissue patches extracted from malignant areas, tissues with low quality, lacking diagnostic value, and devoid of morphological symptoms were removed as done in previous studies[28,30]. Additionally, in the second round of data cleaning, we implemented a criterion to exclude data centers that contributed fewer than 40 slides in total, aiming to mitigate imbalance issues. This exclusion was based on the aggregate number of slides from each center, irrespective of the number of slides per cancer type within each center. The final version includes patches originating from 38 acquisition sites with 29 cancer types. In total, the training set consists of 350,720 patches from 5,824 slides, the test set includes 90,980 patches from 593 slides, and the validation set contains 19,492 patches from 596 slides.

Feature extraction

For the specific purpose of estimating bias in this study, we have conducted a series of experiments and investigated the bias problem from different perspectives. To this effect, we need to extract features from the WSI to perform further processes. For this purpose, we employed two different deep networks. Firstly, we utilized the TCGA features extracted by Riasatian et al.[28], as previously presented in their work. These features were extracted using KimiaNet, a deep neural network based on the DenseNet topology, which underwent fine-tuning, training, and evaluation specifically for the TCGA dataset. The training process of KimiaNet using the TCGA dataset is detailed in Dehkharghanian et al. (2023)[22]. Secondly, we selected EfficientNet as our other baseline model, which was fine-tuned using the selected data samples. In summary, the EfficientNet family of convolutional neural networks employs a composite scaling strategy to deliver cutting-edge performance while being computationally efficient. Its design has found widespread application in computer vision research and practical implementations due to its outstanding performance and resource effectiveness[31]. Consequently, this model represents a more suitable choice for our specific purpose of conducting case studies. However, to facilitate an intensive and in-depth investigation, we deliberately narrowed down our EfficientNet-based analysis to two specific cancer types: Lung Squamous Cell Carcinoma (LUSC) and Lung Adenocarcinoma (LUAD). This selection allowed us to conduct thorough and meaningful case studies toward biased results examination. The reason for using two different network outputs, the pre-prepared features and the next ones extracted by the self-trained network, is to analyze the bias effect on extracted cancerous features at various stages: data distribution, pretraining phase, throughout training, and post-processing steps.

In our experiments, we used a fine-tuned version of the EfficientNet model. The goal was to use the underlying EfficientNet model's capabilities while adapting its behavior to our needs. To achieve this, as shown in Fig. 2,

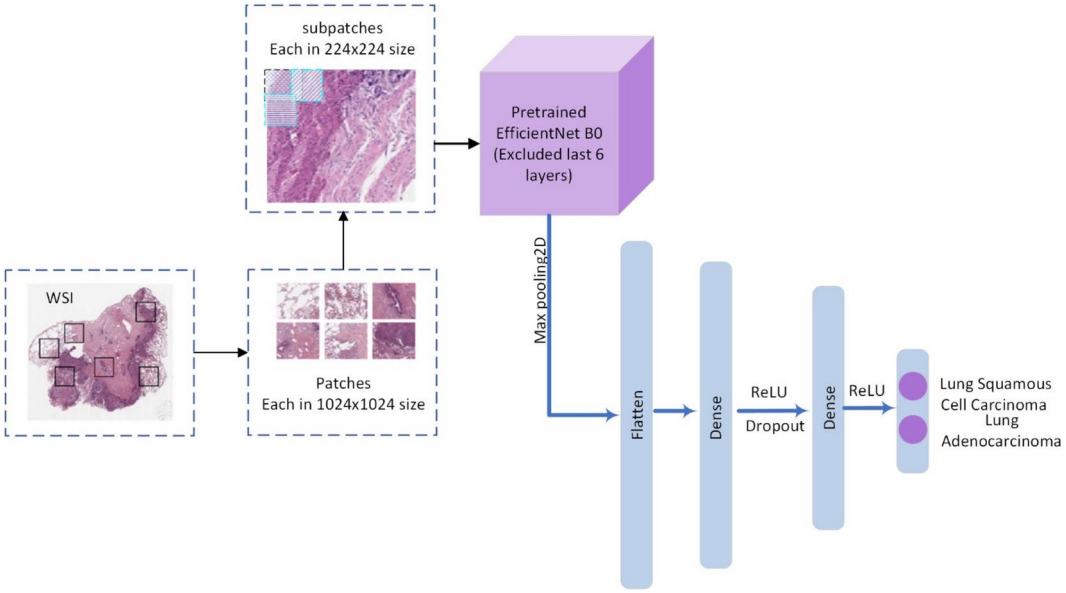


Fig. 2. The figure illustrates the process of adapting input patch sizes for EfficientNet by segmenting patches into smaller subpatches. Then, randomly selected subpatches are processed through a pre-trained EfficientNet architecture. Subsequently, the data passes through the last six layers of the EfficientNet, which are configured to be trainable.

we froze the weights of all layers except for the last six, allowing the model to retain the pre-trained feature extraction benefits while enabling fine-tuning on the top layers to learn task-specific patterns. Subsequent to the frozen layers, we integrated a MaxPooling2D layer with a kernel size of 7x7, followed by a flattening layer. Next, we added a series of densely connected layers that progressively refined the extracted features-beginning with a dense layer containing 1024 units and applying the ReLU activation function along with a dropout layer with a dropout rate of 0.3 to prevent overfitting. Then, an additional dense layer with 256 units was added as the output embedded feature vector, incorporating a kernel regularization technique with an L2 penalty of 0.01, which contributes to controlling the complexity of the model. The final layer of the model is designed for prediction and classification, a dense layer that features two units and employs the softmax activation function.

For the data preparation stage, we employed a sub-patch level procedure that totally differs from the approach utilized by Riasatian et al.[28]. To construct the training datasets, we initially integrated the entire TCGA dataset, encompassing the training, validation, and test subsets associated with both LUSC and LUAD cancer subtypes in order to examine the data distribution effect in the biased training procedure. This step was crucial to ensure that the model does not rely on biased signatures specific to patches from a single slide, as patches from one slide were assigned to only one of the original subsets (training, test, or validation). Also, as part of the preprocessing step[28], patches without cancerous patterns were excluded. This ensures that redistributing samples from one slide across the new training, validation, and test sets does not negatively affect the model's performance in cancer discrimination. Subsequently, we partitioned each original TCGA patch, sized at 1000×1000 pixels, into 25 sub-patches of 224×224 pixels[25], to align with the input size required by EfficientNet-B0, which was originally trained on ImageNet samples. The original patches were split into 25 sub-patches of 224×224 pixels. These sub-patches were created with a step size of 200 pixels, meaning each new patch overlaps the previous one by 24 pixels. This ensures we capture all important cancerous details without resizing the patches. Following this, we proceeded to randomly allocate the prepared data into the training, validation, and test sets with 80%, 10%, and 10% ratios, respectively.

Investigation of potential bias factors

As previously mentioned, KimiaNet is a DenseNet-based model trained on the TCGA dataset with the objective of representing histopathological features. However, research conducted by Dehkharghanian et al.[22] reveals that these extracted features surprisingly possess the capability to discriminate the origin of histopathology patches with a high degree of balanced accuracy. This indicates that the features embed signatures of data origin, irrespective of the model's original purpose of purely learning cancer morphology.

This section provides an in-depth analysis of potential sources of bias that are responsible for embedding data center signatures into the cancer features and may affect the trustworthiness of learning models. Herein, we conduct a series of case studies, each designed to investigate the effect of a hypothetical source of bias on the outcomes of deep learning models. To this end, we apply two deep learning models in our case studies, the pretrained KimiaNet and EfficientNet, to the TCGA dataset.

Case Study 1: Investigating the impact of data sampling

In this section, we investigate the impact of the data sampling process on the performance outcomes of learning models. The objective of this case study is to analyze the distribution of cancer types within the TCGA dataset, focusing on their relationship with data collection centers. Specifically, our objective is to determine whether there is a dependency between cancer types and data collection centers, such as a particular cancer type originating exclusively from specific centers. The presence of such a dependency may lead the learning model to recognize the signature of the data center rather than the distinguishing features of cancerous tissues. To address this concern, we employ a combination of disparity analysis and mutual information (MI) metrics. Our goal is to identify and quantify any existing correlations between cancer types and their respective data collection centers, thus ensuring the reliability of the models developed from this dataset and enhancing the generalizability of the insights obtained.

MI analysis on the TCGA original dataset To provide an analytical description of how cancer types and their origins are correlated, we analyze the fairness of the distribution of cancer types across data sites. For this purpose, we employ disparity analysis using Eq. (1) and MI[32] to assess the dependency between each cancer type and the corresponding data centers and quantify the fairness level in the TCGA data collection process. The full numeric dependency of each cancer/center type is provided in the Supplementary material. The provided table depicts the number of cancers originating from each data center.

$$P(X, Y) = P(X) \cdot P(Y) \tag{1}$$

Disparity Analysis refers to the process of comparing calculated probabilities with observed probabilities to investigate the relationship between two variables, *X* and *Y*, specifically cancer types and data centers within the TCGA dataset. This analysis aims to identify whether these variables are probabilistically independent or if a level of dependence or correlation exists between them. In the context of our case study, we apply this concept to each cancer type included in the TCGA dataset to investigate whether two events (cancer type and data source) are correlated or not. Independence would imply that:

$$P(cancer_i, center_j) = P(cancer_i) \cdot P(center_j) \tag{2}$$

We calculated terms of $P(cancer_i) \cdot P(center_i)$ and $P(cancer_i, center_j)$ for all permutations of cancer types and data centers. However, we observed a disparity between these values, indicating that:

For all *i* and *j*:

$$P(cancer_i, center_j) \neq P(cancer_i) \cdot P(center_j) \tag{3}$$

This emphasizes that the events $cancer_i$ and $center_j$ are not probabilistically independent but rather exhibit a level of dependence or correlation.

After this, to quantify the degree of dependency, we applied the concept of MI using the achieved probability result to gain more insight into the extent of shared site-specific information. MI quantifies how much knowledge two random variables have in common, considering cancer type and data site in our case study. It provides a method to numerically express the statistical reliance or correlation between these variables. As shown in Eq. (4), MI mathematically expresses the uncertainty or randomness for two discrete random variables, corresponding to the cancer type of each patch and its acquisition site.

$$\mathrm{MI}(cancer; center) = \sum_{center} \sum_{cancer} P(cancer, center) \log \left(\frac{P(cancer, center)}{P(cancer)P(center)} \right) \tag{4}$$

where *P*(*cancer*, *center*) is the joint probability of each pair of cancer and center and *P*(*center*) and *P*(*cancer*) are the marginal probabilities.

Correlation visualization using clustered heatmap To enhance data understanding of the dependency of cancer-type distribution across data acquisition sites, we employed a clustered heatmap, considering the correlation metric to cluster distributionally similar data centers. In the context of the TCGA dataset, where certain categories of cancers or classes are significantly more frequent in parts of data centers than others, a clustered heatmap can provide a clear visual summary of the distribution and relationships between these categories. As depicted in Fig. 3, the drawn heatmap, where the x-axis represents data centers and the y-axis represents cancer types, visualizes a comprehensive overview of the prevalence of each cancer type within specific data centers. In Fig. 3, "prevalence" refers to the proportion of a specific cancer type within a given data center. Prevalence was calculated as the number of samples for a specific cancer type (e.g., cancer type X) coming from a center (e.g., Center A), divided by the total number of samples from that center. This gives an indication of how prevalent a particular cancer type is within each data center. The exact numbers of samples are detailed in the Supplementary material. The clustered heatmap employed these values to visualize data distribution.

In other words, $cell_{(i,j)}$ in the clustered heatmap indicates the dependency intensity of $cancer_i$ on $center_j$. This intensity is calculated as the ratio of the number of $cancer_i$ samples that originate from $center_j$ to the total number of $cancer_i$ samples. The final mapping represents disparities between clusters or distributions of cancerous samples over data centers. In addition, the heatmap incorporates clustering based on a correlation metric to group together data centers that are distributionally similar. This means that data centers with similar profiles in terms of the types of cancer are grouped together in the visualization. For example, as shown in Fig. 3,

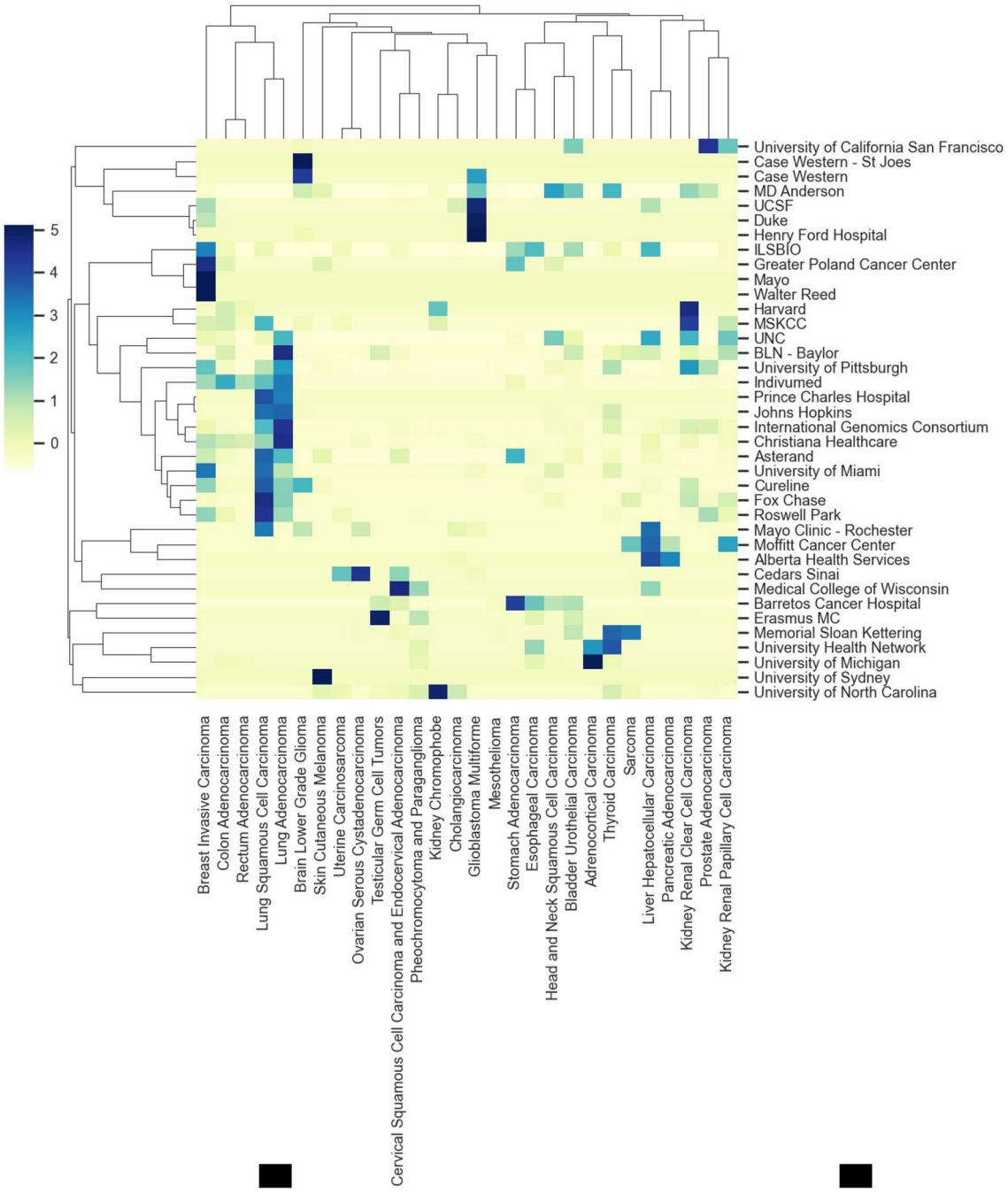


Fig. 3. Unveiling Imbalance in Cancer-Type Distribution Across Data Centers using Clustered Heatmap Analysis: This clustered heatmap visually depicts the varying prevalence of cancer types in different data centers. The color intensity reflects prevalence magnitude, revealing distinct patterns and disparities among clusters.

Lung Adenocarcinoma samples mostly originate from the International Genomics Consortium and Christiana Healthcare. Then, the corresponding cells' color represents higher dependency values for these two data centers. On the other hand, this cancer type has nearly no samples from the ILSBIO and MD Anderson centers, as the color in the visualized heatmap is close to white. So, this unfair data sample distribution can lead to capturing the International Genomics Consortium and Christiana Healthcare centers' signatures within the training procedure.

MI analysis on the TCGA DataSet after correlation removing In this case study, we tested the validity of our hypothesis by creating an uncorrelated condition within a subset of the TCGA dataset. The MI values of zero, after artificially removing correlations in the controlled subset, and 1.47 for the original dataset clearly indicate the absence and presence of dependency, respectively. These results underline significant disparities in cancer-type

distribution across data centers, emphasizing the need to consider these factors to ensure the fairness and reliability of predictive models developed from this dataset. Hence, the use of disparity analysis and MI has provided a quantitative understanding of this dependency, underscoring the importance of considering these factors in the fairness of data collection processes.

Case Study 2: Investigating the impact of patching
The primary aim of this research is to assess the impact of the patching process on biased results in the classification of cancer types and the identification of data source site institutions. This involves an analysis of how the preparation process of TCGA data, particularly the division of whole slide images (WSIs) into patches, influences the outcomes of our deep learning models, KimiaNet and EfficientNet.

Patching process investigation on KimiaNet preprocessing After completing the data-gathering process, which includes two cancer types originating from three data centers, we undertake an additional step to analyze the preparation process of the collected TCGA data before feeding it into KimiaNet. As explained in previous sections, TCGA consists of WSIs, each of which includes multiple cropped sections known as patches. Due to the exceedingly large size of WSIs and computational limitations, the model required patch data as input after undergoing preprocessing steps. In the referenced research paper[28], one of KimiaNet's preprocessing steps involved dividing the WSIs into three sets: training, testing, and validation. Following this, the slides into each set were divided into patches. In other words, the network processes patches from individual slides collectively. This patching process raised our concerns about the potential impact of this step on biased results, which is cheating through co-slide patches (i.e., patches from the same slide) in the center classification process.

To determine whether this patching step affects the final biased results, we conduct three independent test cases by applying the *k*-Nearest Neighbor (*k*-NN) classifier to cancer features which is extracted by deep models with the goal of discriminating between cancer types and identifying data source site institutions. In addition, we examine the significance of biased effects caused by co-slide patches by considering both their inclusion and exclusion when finding the nearest neighbors in *k*-NN. The following experiments were conducted using a training set consisting of 350,720 patches from 5,824 slides, while the test set includes 90,980 patches from 593 slides.

1. **Test Patches Search in Test Set.** As the first test case, a leave-one-out searching approach is used for the *k*-NN classifier, where the nearest neighbors of each test set's sample are determined from the entire test set after leaving out the candidate sample (test-in-test). As depicted in Table 1, we achieved 99% and 96% balanced accuracy values for cancer and center determination types, respectively. It is important to note that using the test set as the search space for the KNN classifier may not be practical in real-world scenarios. This is because, in real-time applications, we wouldn't typically have access to all future test data in advance to consider as the search space. To classify each test sample effectively using this method, we would need to wait for additional data to be collected and added to the test set. However, as more data is accumulated over time, this approach might become more feasible. Test Patches Search in Training Set. In the second test case, the training set is used as the search space for test samples (test-in-train) to find nearest neighbors. Surprisingly, the results demonstrate a significant disparity between this and the former test case, with the balanced center accuracy dropping from 96% to 62% and the cancer balanced accuracy dropping from 99% to 66%. Since the training and test samples were collected from equal data centers, the reduction in balanced accuracy values could not be just due to the disparities between training and test set distribution.
2. **Test Patches Search in Test Set with Co-slide Exclusion.** Afterward, we conduct an additional experiment using a restricted form of *k*-NN (R*k*-NN) to investigate the cause behind this decline in balanced accuracy. In this particular test, mirroring the first one, we repeated the *k*-NN classification process by identifying neighbors for each test sample within the test set. However, this time, we excluded the co-slide patches for each candidate patch when searching for its nearest neighbors. We aim to investigate the influence of co-slide patches on the final biased results. The classification outcomes for both tasks, as seen in Table 1, demonstrate the discrimination of acquisition sites with a balanced accuracy of 48% and cancer types with a balanced accuracy of 66%. Upon comparing the results achieved in the aforementioned test cases, the test-in-train search result (second test) is closer to the outcome of the R*k*-NN classifier (third test). However, both of these results significantly differ from the original test-in-test results (first test). A shared characteristic between the

Test cases	Cancer-Acc (%)	Center-Acc (%)
KimiaNet Patching: Test-in-Test	99	96
KimiaNet Patching: Test-in-Train	66	62
KimiaNet Patching: Test-in-Test with exclusion condition	66	**48**
EfficientNet Patching: Test-in-Test	**83**	68
EfficientNet Patching: Test-in-Train	**83**	68
EfficientNet Patching: Test-in-Test with exclusion condition	82	67

Table 1. Investigation of patching bias through three test cases for cancer type classification balanced accuracy (Cancer-Acc) and data origin balanced center accuracy (Center-Acc). The achieved results represent similar results for first and third test cases, evidence for contributing patching step in biased results. Significant values are in bold.

second and third experiments, not present in the first one, is the exclusion of all patches from a specific WSI within the search space. Of note, the original TCGA dataset was separated into training and test sets at the slide level, meaning that patches from the same slide are included either in the test or training set.

The obtained results from these experiments show that the classifier's performance is highly dependent on the inclusion of co-slide patches in the comparison pool. The initial high balanced accuracy in the test-in-test approach likely reflects an overestimation of the true classification power, potentially due to the test samples having very similar or identical counterparts within the test set. The drop in balanced accuracy when co-slide patches are excluded indicates that the classifier may be recognizing slide-specific features rather than generalizable patterns for cancer and center identification.

Patching process investigation on EfficientNet preprocessing To explore this hypothesis from an alternative perspective, we investigate its validity by randomly selecting patches from four distinct data centers-namely, Johns Hopkins, Asterand, Indivumed, and Roswell Park in a fully balanced manner. Subsequently, we utilize the customized EfficientNet to extract the corresponding features. Following this, we employ the *k*-NN algorithm aimed at the test-in-test search, both with and without the exclusion of co-slide patches and test-in-train search. The results obtained from these experiments consistently demonstrated a balanced accuracy of about 83% for discriminating between cancer types and a balanced accuracy of 68% for distinguishing between center types. As previously stated, in contrast to the KimiaNet preprocessing procedure, our preprocessing for EfficientNet involved sub-patch level analysis, treating each sub-patch as an independent entity without requiring the presence of its corresponding co-slide patches. Consequently, the patch selection strategy for the Efficientnet training does not cause a reduction in balanced accuracy resulting from the exclusion of co-slide features, which may arise due to slide-level biased training of KimiaNet, as previously mentioned. This presents the impact of patch selection on model performance, suggesting that patch-level biases can significantly affect classification accuracy. Thereby, this behavior confirms the existence of a type of shortcut into the patches corresponding to the same slide. For example, two patches belonging to different slides but with the same center and cancer type have dissimilar features compared to their co-slide patches. These shared features create a discernible pattern or shortcut, indicating the existence of a slide-specific signature within each patch.

One contributing factor would be specific texture patterns or color distributions that are unique to each slide. Variations in staining techniques or tissue properties could contribute to distinct texture and color characteristics within patches from the same slide. For instance, patches from the same slide may have different textures and colors due to differences in staining methods or tissue features. In addition, biological variations within different slides could lead to unique features in patches. These variances might comprise genetic differences, tumor heterogeneity, or variations in tissue composition.

In conclusion, our findings underscore the critical role of the data preparation process in the performance and bias behavior of deep learning models. Specifically, the way patches are prepared and utilized in training can introduce or mitigate biases. The patching process, particularly the treatment of co-slide patches, has a pronounced impact on model accuracy and bias. In the "test-in-test" scenario in the KimiaNet experiments, when the classifier searches for the nearest neighbors of a patch feature, the search space includes co-slide patches within the test set. As a result, the classifier may rely on uncancerous signatures specific to the slide, which can act as shortcuts to classify the test samples' data center labels. In contrast, in the "test-in-train" scenario, since the classifier searches within the training set (which does not contain co-slide patches from the test set), it cannot use these slide-specific signatures to assign data center labels. On the other hand, this condition does not apply to the dataset used for EfficientNet, which is why we do not observe the gap between the results of the "test-in-test" and "test-in-train" experiments. This research confirms the existence of patch selection biases, challenging us to refine our data preparation techniques to ensure more unbiased and accurate outcomes in cancer detection and classification tasks. In addition, classifiers with a similar design to *k*-NN, which rely on the selection of nearest neighbors, could potentially benefit from this exclusion method. By eliminating closely related data points in the training phase, these classifiers may also achieve a more realistic estimation of their predictive ability. This strategy can be especially beneficial in scenarios where the objective is to detect patterns that are not specific to the dataset's specific structure but rather indicative of broader trends that would apply to unseen data.

Building on these findings, we conducted an additional experiment to investigate whether specific patterns were embedded in the suspicious biased patches of a single slide. For this purpose, we mapped these bias patches back onto the slide to analyze potential spatial patterns but did not find any consistent fixed points or identifiable patterns within the patches. This observation led us to hypothesize that the color factor, consistent across the slide and present in all patches, might be a contributing pattern. This hypothesis aligns with the possibility that unique staining procedures in each data center could introduce color-based biases. This realization motivated the Stain Effect Investigation in Case Study 4, which explores the potential impact of staining variability on model performance and bias.

Case Study 3: Existence of site-specific patterns in cancer-based features

The reduction in balanced accuracy resulting from the exclusion of co-slide patches during the nearest neighbor determination step motivated us to conduct a deeper analysis of the results obtained by R*k*-NN in the previous section. Our aim is to explore additional sources of bias.

During this analysis phase, we carefully selected patches whose origin center was correctly predicted in the third test and identified their nearest neighbors within the classification procedure. Interestingly, our findings revealed that a significant 81% of these voters (i.e., most similar patches) were associated with the same cancer type as the query patch. This indicates that patches from the same cancer and center types share more similar features than those from the same cancer type but with different origin centers.

This observation indicates the existence of site-specific patterns among the features of each cancer type. To provide further clarification, in relatively unbiased deep models, a feature (F_1) associated with cancer of C from center of A should not have exhibited significant differences from a feature (F_2) associated with same cancer (C) but different center (B). This is also should be valid with a feature from the same cancer in the same center. However, contrary to this expectation, differences between F_1 and F_2 have been observed, leading to a strong correlation between cancerous features of a specific data center. For example, each data center might have specific protocols for data acquisition, including imaging techniques, equipment calibration, and image resolution. If these protocols are consistent within a data center but vary across centers, the data from the same center would naturally be more similar to each other.

Cancer-specific analysis of Rk-NN voters Subsequent to the previous section, we go one step deeper into analyzing the nominated nearest neighbors of correctly classified patches' origin. In other words, when the data center classifier applies to cancer features, the center labels are chosen based on the closest features' label to the candidate feature. To do so, we extend our analysis by independently examining the number of samples from specific cancer types that were chosen as the nearest neighbors to examine whether specific cancer types may include the data center signature. For this purpose, we computed the frequency rate relative to the total number of data samples within the test set, as shown in Table 2, representing the scale of each cancer type's collaboration in the biased outcome of the model. The ordered ratio of data cancer types in Table 2 reveals heterogeneous contributions within cancer types. Following the ratio ordering, we categorized the patches into two groups based on frequency rates, with values either more or less than 1.6. This categorization allows for separate discrimination of data origin based on the contributions of cancer type. The feature classification results, summarized in Table 3, confirm our assumption of significantly biased accuracy for patches with contribution rates exceeding 1.6, achieving 59% and 91% balanced accuracy rates for center and cancer classification tasks, respectively. Conversely, the test case for the other patch category, with a contribution rate of less than 1.6, led to significant reductions in accuracy rates: 41% and 63%, respectively. Despite the larger size of the dataset in the category with the lower contribution rate, the accuracy reduction could not be attributed to a decrease in dataset size. Therefore, we can conclude that certain cancer types might have more distinctive morphological features that make it easier for the model to learn and recognize than others. If these features are both unique and prevalent in the training data, the model will more effectively identify and classify these cancer types, leading to a bias in performance

Cancer type	#Votes	# Test Samples	Ratio
Kidney Renal Clear Cell Carcinoma	24789	10071	**2.46**
Uterine Carcinosarcoma	1980	829	**2.39**
Sarcoma	7327	3227	**2.27**
Testicular Germ Cell Tumors	1629	726	**2.24**
Ovarian Serous Cystadenocarcinoma	3792	1747	**2.17**
Kidney Chromophobe	4321	2064	**2.1**
Prostate Adenocarcinoma	9464	4799	**1.97**
Head and Neck Squamous Cell Carcinoma	5570	3242	**1.72**
Pancreatic Adenocarcinoma	1786	1051	**1.7**
Stomach Adenocarcinoma	6887	4299	**1.6**
Breast Invasive Carcinoma	15045	10297	1.46
Glioblastoma Multiforme	4524	3248	1.4
Liver Hepatocellular Carcinoma	6625	5171	1.28
Skin Cutaneous Melanoma	3410	2688	1.27
Kidney Renal Papillary Cell Carcinoma	4013	3249	1.23
Thyroid Carcinoma	10286	8644	1.19
Cholangiocarcinoma	1913	1618	1.18
Lung Squamous Cell Carcinoma	4560	3869	1.18
Lung Adenocarcinoma	4123	3979	1.03
Colon Adenocarcinoma	3505	3739	0.94
Adrenocortical Carcinoma	1234	1382	0.89
Esophageal Carcinoma	1446	1648	0.88
Brain Lower Grade Glioma	2309	2733	0.85
Bladder Urothelial Carcinoma	1507	1852	0.81
Rectum Adenocarcinoma	1004	1714	0.58
Cervical Squamous Cell Carcinoma	493	1286	0.38
Pheochromocytoma and Paraganglioma	562	1797	0.31
Mesothelioma	0	11	0

Table 2. Contribution rate of each cancer type within data center discrimination. Significant values are in bold.

Test set	Cancer Acc (%)	Center Acc (%)	Diff. Rate (Cancer)	Diff. Rate (Center)
Baseline	66	48	–	–
< 1.6 frequency rate	91	59	0.38	0.22
> 1.6 frequency rate	63	41	0.04	0.14

Table 3. Comparative analysis of model accuracy across different cancer types based on contribution rates. This table presents the accuracy rates achieved for center and cancer classification tasks, categorizing the test cases into two distinct groups according to their frequency rates: those with contribution rates exceeding 1.6 and those below 1.6. The accuracy percentages demonstrate the significant impact of contribution rates on model performance, highlighting the disparity in accuracy between cancer types with higher versus lower representation in the dataset. To compare the results of these experiments with the baseline, we calculate the ratio of the difference between the target experiment value and the baseline value to the baseline value itself.

	Cancer_num	Center_num	Cancer-Acc (%)	Center-Acc (%)
Fair subset	2	2	79	66
Correlated subset	2	2	95	95

Table 4. Obtained results on balanced and imbalanced cases.

favoring these cancers. Although accuracy changes were observed for both tasks, the center accuracy change was more pronounced than the cancer-related change.

Case Study 4: Investigation of bias in efficientnet features

In this section, we step back to explore sources of bias by designing experiments that manipulate training procedures with raw samples. As KimiaNet is a pre-trained network on WSIs, we did not have control over bias investigation during its training; therefore, we conducted the experiments using extracted features. However, for this phase of the experiments, we fine-tuned EfficientNet to extract features indicative of cancer. To this end, we created two datasets: the first under completely fair conditions, with no correlation between cancer types and their data origins, and the second with an intentional correlation between cancer types and data origins.

The goal of this experiment is to compare the results achieved when training the model on this artificially correlated dataset with those obtained using a fully balanced dataset, where both cancer types are evenly distributed across the two data centers. This comparison allows us to examine how data center-specific patterns embedded in the training samples affect the deep model learning process.

For the first dataset, we randomly selected two distinct data acquisition sites, namely Asterand and Prince Charles Hospital. From each site, we collected a total of 200 patches, resulting in 5,000 sub-patches, equally divided between the two cancer types. For the second dataset, we intentionally created a correlation between cancer types and their center of origin. We maintained the same data sample size as the first dataset but selected each cancer type from only one specific data center. Both datasets then underwent a partitioning process: 10% of the data was allocated for validation purposes, an additional 10% was set aside for our dedicated test set, and the remaining 80% formed the training set, which was used to train the EfficientNet model. Next, EfficientNet was applied to each dataset separately to compare their outcomes. After completing the EfficientNet training step, we extracted the cancerous features obtained from the deep model training. Subsequently, we utilized the k-NN classifier to discriminate data acquisition sites and cancer types. We repeated this procedure for each data subset separately. The results of classification steps are presented in Table 4.

The discrimination results of the first data subset represent 79% balanced accuracy for cancer balanced accuracy and 66% balanced accuracy for data center discrimination. On the other hand, the results for the subset samples with intentional correlation show surprisingly completely different results, 95% balanced accuracy for both cancer and data center discrimination. The variation in cancer classification balanced accuracy between the two datasets-79% for the first dataset and 95% for the second-can be attributed to the difference in how the data was collected and the presence of dataset bias. In the first dataset, both cancer types were collected equally from two data centers. This method of data collection minimizes bias related to the data center of origin, as both cancer types are represented equally across the two centers. The model's task is to learn the intrinsic features that differentiate LUAD and LUSC, irrespective of their origin, making the classification task purely based on cancer-specific features. Hence, the lower balanced accuracy in the first dataset suggests that the model has a more challenging task as it must rely solely on the cancerous features for classification without the help of bias. The absence of a simple pattern or correlation forces the model to focus on the more nuanced differences between LUAD and LUSC, which is a more difficult task and might not be as accurately performed depending on the complexity of the cancer features and the model's capacity.

In the second dataset, a clear correlation is established between cancer types and data centers (LUSC exclusively from Asterand and LUAD exclusively from Prince Charles Hospital). This introduces a strong bias because the model might not only learn the distinguishing features of the cancer types themselves but also pick up on any differences in the data collection process, instrumentation, or other characteristics specific to each data center. The model could leverage this additional information (bias) to achieve higher balanced accuracy,

not necessarily because it has become better at recognizing cancer-specific features but because it has learned to associate certain features with the data centers.

From the data center's achieved balanced accuracy values perspective, the lower accuracy (66%) in classifying the center of origin from the first dataset likely comes from the balanced nature of data collection. Since both cancer types were equally collected from both data centers, any unique center-specific features are evenly distributed across both cancer types. This distribution makes it more challenging for the *k*-NN classifier to distinguish the center of origin based solely on the cancerous features, as these features are not uniquely tied to one center or the other. On the other hand, the significantly higher balanced accuracy (95%) observed with the second dataset can be attributed to the clear correlation between cancer types and their respective centers of origin. In this scenario, any center-specific characteristics (e.g., variations due to data collection methods, instrumentation, or processing) inadvertently become markers that the *k*-NN classifier can use to determine the center of origin. Essentially, the model may not identify the center based on the cancerous features themselves but rather leverage the inherent bias introduced by collecting each cancer type from a specific center.

The experimental results from using deep models and *k*-NN classifiers to identify cancer types and their centers of origin across two datasets reveal significant insights into the impact of dataset bias on machine learning accuracy. In both tasks-cancer type classification and center of origin discrimination-the second dataset, characterized by an intentional correlation between cancer types and specific data centers, achieved markedly higher accuracies (95%) compared to the first, balanced dataset (79% for cancer type classification and 66% for center of origin). These findings represent the high influence of dataset composition on model performance. While the high balanced accuracy in the biased dataset may initially cause advantages, it actually highlights reliance on dataset-specific biases rather than an intrinsic understanding of the desired task. This reliance can lead to models that perform well under specific, biased conditions but may fail to generalize to broader, more diverse scenarios.

Stain effect investigation Histopathology images are typically stained to highlight specific cellular structures and tissue components. Variations in staining techniques used by different laboratories and facilities may result in varied color intensity and distribution. This diversity in staining might result in inconsistencies in how features are represented in histopathology images, affecting the accuracy and generalizability of models trained on such data[33,34]. To have an in-depth analysis, we assess the influence of color pixels by adding random noise to RGB channels separately and then transforming the dataset into grayscale, and subsequently reproducing the classification procedure, which we named this process Noise-Based Grayscale Normalization (NBGN). The purpose of converting images to grayscale and adding noise was not to address bias but also to explore the effect of color on data center accuracy. Our goal was to investigate how the RGB channels contribute to center-specific patterns. In this method, we recognized that converting the images from color to grayscale might still transfer the same center-specific patterns from the RGB channels to grayscale. Therefore, we introduced random noise to disrupt these patterns. By doing this, we wanted to explore how much the RGB color channels contribute to creating patterns specific to each data center.

To implement this idea, as shown in (5), we injected random noise into each RGB channel of every patch to mitigate the influence of site-specific stains and then retrained EfficientNet using the grayscale dataset. As presented in Table 5, we explored several thresholds for noise injection, aiming to have minimal disruption to cellular structures and tissue components. In other words, first, we used an RGB training set to train the efficientNet and classify both cancer and center samples of an RGB test set. The obtained results are presented in the first column of Table 5. Next, we converted the same RGB training and test sets to a grayscale version and retrained the efficientNet using the gray training set, and assessed the center and cancer accuracy using the gray test set features. We repeated these experiments for diverse threshold values, as shown in Table 5, to find the optimal one. The threshold controls the level or intensity of random noise added to the RGB channels of image patches to reduce the impact of site-specific stains without significantly distorting important cellular and tissue structures. The noise injection aims to make a model like EfficientNet focus more on structural features rather than color variations due to staining, enhancing its ability to discriminate cancer accurately. The optimal threshold balances minimizing site-specific stain influences while preserving crucial image details for cancer detection. This approach seeks to mitigate potential negative effects on the results of cancer discrimination. Compared to the original RGB dataset, all thresholds result in a higher drop rate in center discrimination outcomes, even when not considering their comparatively lower influence on cancerous structures. This underscores the notable significance of site-specific stains as learned shortcuts. However, the optimal threshold value would lie between 0.75 and 1, as indicated by the findings. Samples of converted patterns from RGB to gray using 0.75<TH<1 are represented in Fig. 4.

	RGB version (%)	Th < 0.75 (%)	Th < 0.5 (%)	Th < 0.4 (%)	Th < 0.6 (%)	0.75 < Th < 1 (%)
Cancer-Acc	84	66	75	72	77	**82**
Center-Acc	75	43	54	53	55	**65**

Table 5. Noise-based grayscale normalization results The table presents the varying levels of random noise applied to each RGB channel to reduce the effects of site-specific staining variations. Significant values are in bold.

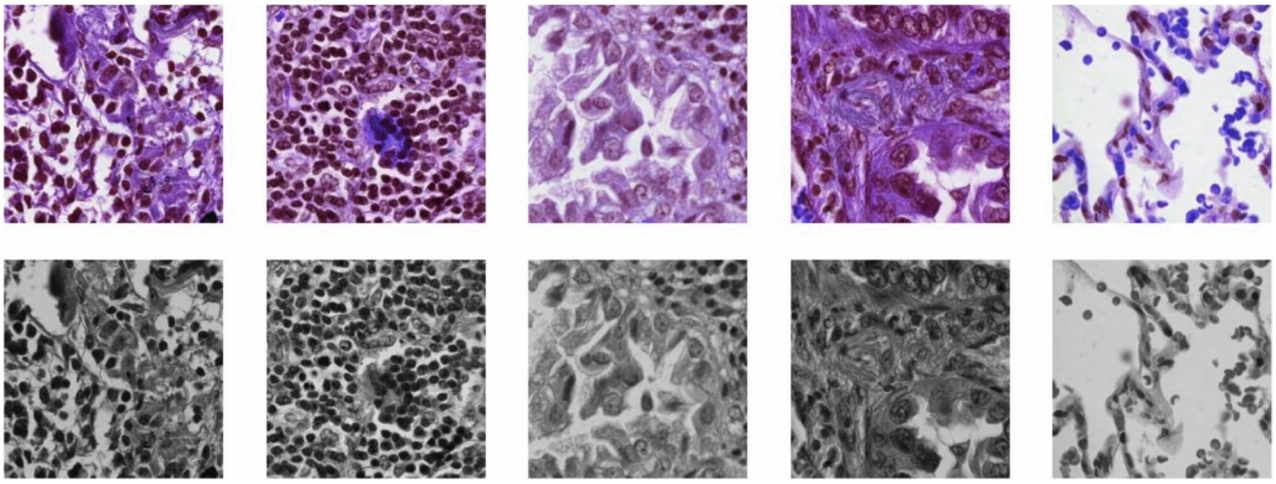


Fig. 4. The figure displays histopathology samples alongside their converted versions. The first row illustrates the original RGB samples, while the second row shows the corresponding grayscale versions, which have been converted using the process described, with a threshold (TH) range of 0.75 to 1.

Training	Cancer-Acc		Center-Acc	
Samples	RGB test	NBGN test	RGB test	NBGN test
RGB	77%	–	59%	–
NBGN	–	74%	–	50%

Table 6. Stain effect investigation. Comparative analysis of cancer detection and data center discrimination accuracies using RGB and samples after applying NBGN. The table summarizes the results of two experiments: the first using RGB images for training and testing with EfficientNet, and the second using normalized images converted from the RGB training samples. Accuracies are presented for cancer type identification and data center origin determination for each image set.

$$\begin{aligned}\mathrm{Image}_{\mathrm{out}}(x,y) = \mathrm{repeat}(\mathrm{expand}(\mathrm{convert\_to\_grayscale}(&\\ \mathrm{clip}(B(x,y)\times \mathrm{random}(a,b),0,255),&\\ \mathrm{clip}(G(x,y)\times \mathrm{random}(a,b),0,255),&\\ \mathrm{clip}(R(x,y)\times \mathrm{random}(a,b),0,255))),3)&\end{aligned} \tag{5}$$

where

- $\mathrm{clip}(C \times C_{\mathrm{inj}}, a, b)$ ensures the modified pixel values of channel C (where $C \in \{B, G, R\}$) remain within the valid intensity bounds.
- $\mathrm{convert\_to\_grayscale}()$ converts the RGB image with modified channels into a grayscale image.
- $\mathrm{expand}()$ and $\mathrm{repeat}()$ manipulate the dimensionality of the grayscale image to mimic that of a three-channel RGB image.
- $\mathrm{random}(a, b)$ represents the generation of a random injection factor between a and b for each channel.

To investigate the impact of stain variations from a different perspective, we established an optimal threshold range (0.75<Th<1) for our experiments assessing how RGB stains influence the features learned by deep models. In the first experiment, we compiled training data from three distinct data centers, training EfficientNet on this RGB dataset to extract cancerous features. Subsequently, we utilized this trained model to extract cancerous features of an RGB test set, employing a *k*-NN classifier to assess cancer type and center origin accuracy. The results, detailed in Table 6, showed discrimination accuracies of 77% for cancer and 59% for data center identification from the RGB samples.

For the second experiment, we applied the same optimal threshold to convert the initial experiment's training samples from RGB to grayscale using NBGN method before training EfficientNet. This model was then used to evaluate normalized test samples, with the normalized training data serving as the search space for classification. Presented in Table 6, the findings for the grayscale test set revealed cancer detection and data center discrimination accuracies of 74% and 50%, respectively.

The decrease in accuracy for both tasks was calculated as a relative change, rather than a simple subtraction, to provide a more meaningful measure of the impact. Relative change shows how much the accuracy decreased

compared to the baseline (RGB accuracy in this case), offering better insight into the effect of applying NBGN method. The calculation for relative decrease is as follows:

$$\text{Relative Decrease} = \frac{\text{RGB Accuracy} - \text{NBGN Accuracy}}{\text{RGB Accuracy}} \tag{6}$$

Comparing the results from both experiments reveals that using normalized images for training, which eliminates color information-a potential source of bias-leads to a decrease in accuracy for both cancer detection and data center identification in the second experiment. Specifically, the decrease in cancer detection accuracy is by 4%, while the decrease in data center identification accuracy is more noticeable at 15%. The decrease in accuracy, especially more pronounced in data center identification (a reduction of 15%), indicates that color information might introduce a bias that aids in the discrimination process, as an unrelated feature in learning cancerous feature. By removing this potential source of bias (color), the model's generalization capability is tested, revealing a dependency on color for achieving higher accuracy levels.

Comparison with other stain normalization methods To evaluate the effectiveness of our stain normalization method, we compared it with results achieved by two widely used techniques: Reinhard normalization and Optical Density (OD) transformation. After applying each normalization method, the data samples were converted to grayscale, following the same steps as in our method. The deep model was then trained using the normalized data to extract cancerous features for further analysis.

Optical Density transformation is a method based on the Beer-Lambert law, which relates how much light is absorbed by a staining agent to its concentration[35]. This technique transforms pixel intensities into optical density values using the formula:

$$OD = -\log\left(\frac{I}{I_0}\right) \tag{7}$$

Here, I is the intensity of a pixel, and I_0 is the background light. The OD transformation helps separate the effects of different stains in an image, making it easier to standardize the staining patterns.

Reinhard normalization is a simple yet effective stain normalization technique widely used in digital pathology. It works by standardizing the color distribution of an image to match a reference distribution, typically defined by predefined mean and standard deviation values for each color channel. This method aligns the image's color distribution to a common reference, reducing staining variability across different samples[36]. The method involves the following steps:

1. Convert the image to the LAB color space, which separates luminance (L) and chromaticity (A and B) channels.
2. For each channel (L, A, and B), the mean and standard deviation are adjusted to match the target reference values:

$$I' = \frac{(I - \mu)}{\sigma} \cdot \sigma_r + \mu_r \tag{8}$$

where

- I is the pixel intensity.
- μ and σ are the mean and standard deviation of the image channel.
- μ_r and σ_r are the target reference mean and standard deviation.

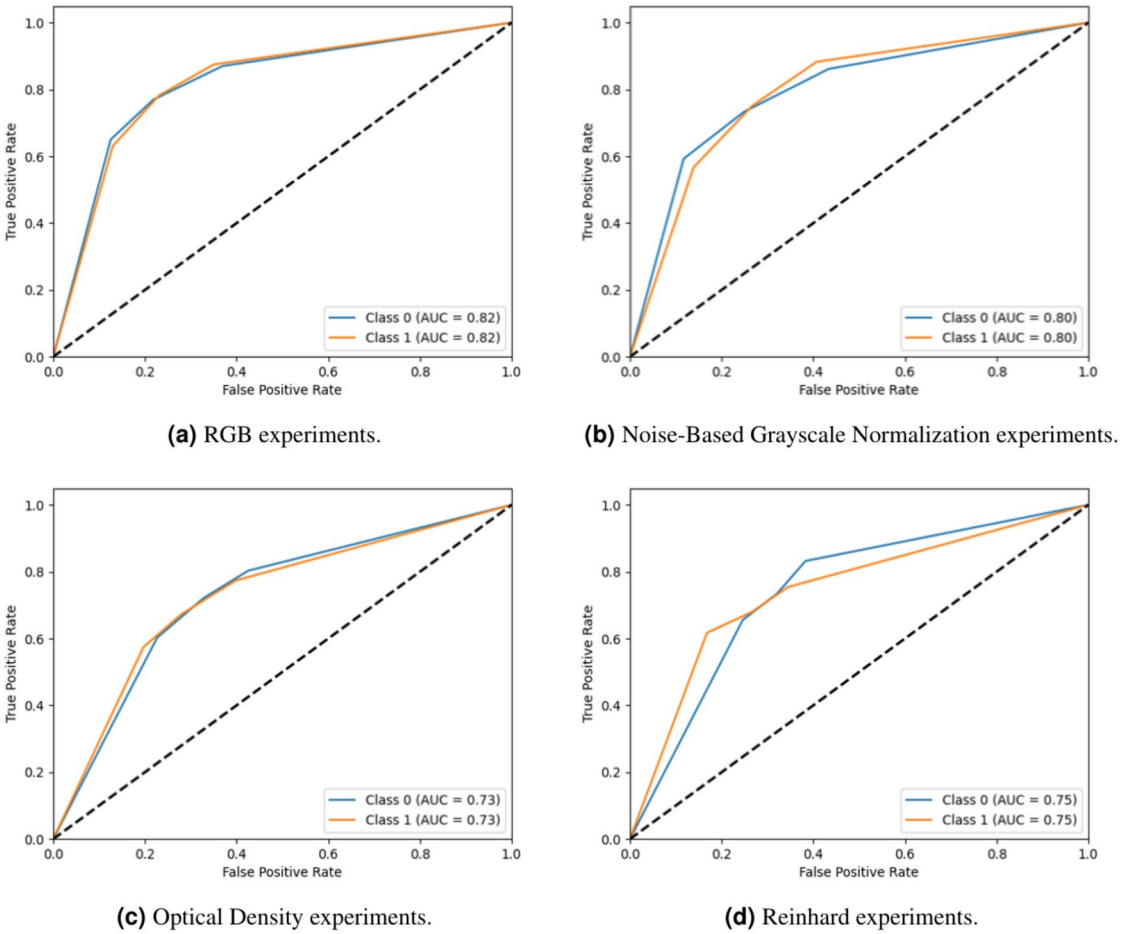
3. Convert the normalized LAB image back to the RGB color space.

To determine the target reference values (μ and σ) for normalization, we calculated the global mean and standard deviation of the LAB channels from a representative subset of the dataset. Specifically, each image in the subset was converted to the LAB color space, and the mean and standard deviation for the L, A, and B channels were computed. These per-image statistics were then aggregated by averaging across all images in the subset to derive the global reference values. In the following, we applied these two stain normalization methods to our dataset samples. The primary goal was to evaluate how effectively these methods neutralized the unique center-based color impacts compared to the Random Noise Injection method. After normalization, the deep learning model was retrained using the normalized datasets to extract cancer-specific features. Notably, the initial weights were kept consistent across all model training processes for the different normalization methods. The extracted features were then evaluated using a KNN classifier to classify the samples based on their centers.

Based on the results obtained, as shown in Table 7, when the NBGN transformation was applied, the cancer detection accuracy dropped slightly to 74%, while the center identification accuracy decreased to 50%, suggesting that the NBGN method effectively reduced center-specific biases while maintaining competitive cancer detection performance. In comparison, the OD transformation achieved a cancer detection accuracy of 66%, but the center identification accuracy remained relatively high at 61%, indicating that OD transformation was less effective in neutralizing center-specific effects. The Reinhard normalization method performed well, achieving a cancer detection accuracy of 71% while significantly reducing the center identification accuracy to 48%. This demonstrates that Reinhard normalization effectively mitigates center-specific biases while maintaining good

	Cancer_Acc (%)	Center_Acc (%)
RGB version	77	59
NBGN transformation	74	50
OD transformation	66	61
Reinhard transformation	71	48

Table 7. Performance comparison of cancer detection accuracy and center identification accuracy across different stain normalization methods. The table highlights the trade-offs between reducing center-based biases and maintaining cancer detection accuracy for RGB, Noise-Based Grayscale Normalization, Optical Density, and Reinhard transformations.



(a) RGB experiments. **(b)** Noise-Based Grayscale Normalization experiments.

(c) Optical Density experiments. **(d)** Reinhard experiments.

Fig. 5. ROC curves for cancer classification under different experimental conditions: RGB (**a**), Noise-Based Grayscale Normalization (**b**), Optical Density (**c**), and Reinhard (**d**). The curves illustrate the model's performance in each experiment, with the AUC values providing a comparative measure of discrimination ability for cancer detection.

cancer detection accuracy compared to original results obtained from original RGB samples. These results highlight that the NBGN and Reinhard methods demonstrated strong potential for unbiased analysis in digital pathology, with NBGN achieving slightly higher cancer detection accuracy and Reinhard effectively balancing bias mitigation with cancer detection performance.

To further analyze the performance of our model and the impact of color information, we extracted the ROC curves and calculated the Area Under the ROC Curve (AUROC) for both cancer classification and data center identification under RGB and normalized conditions using NBGN, OD and Reinhard transformations.

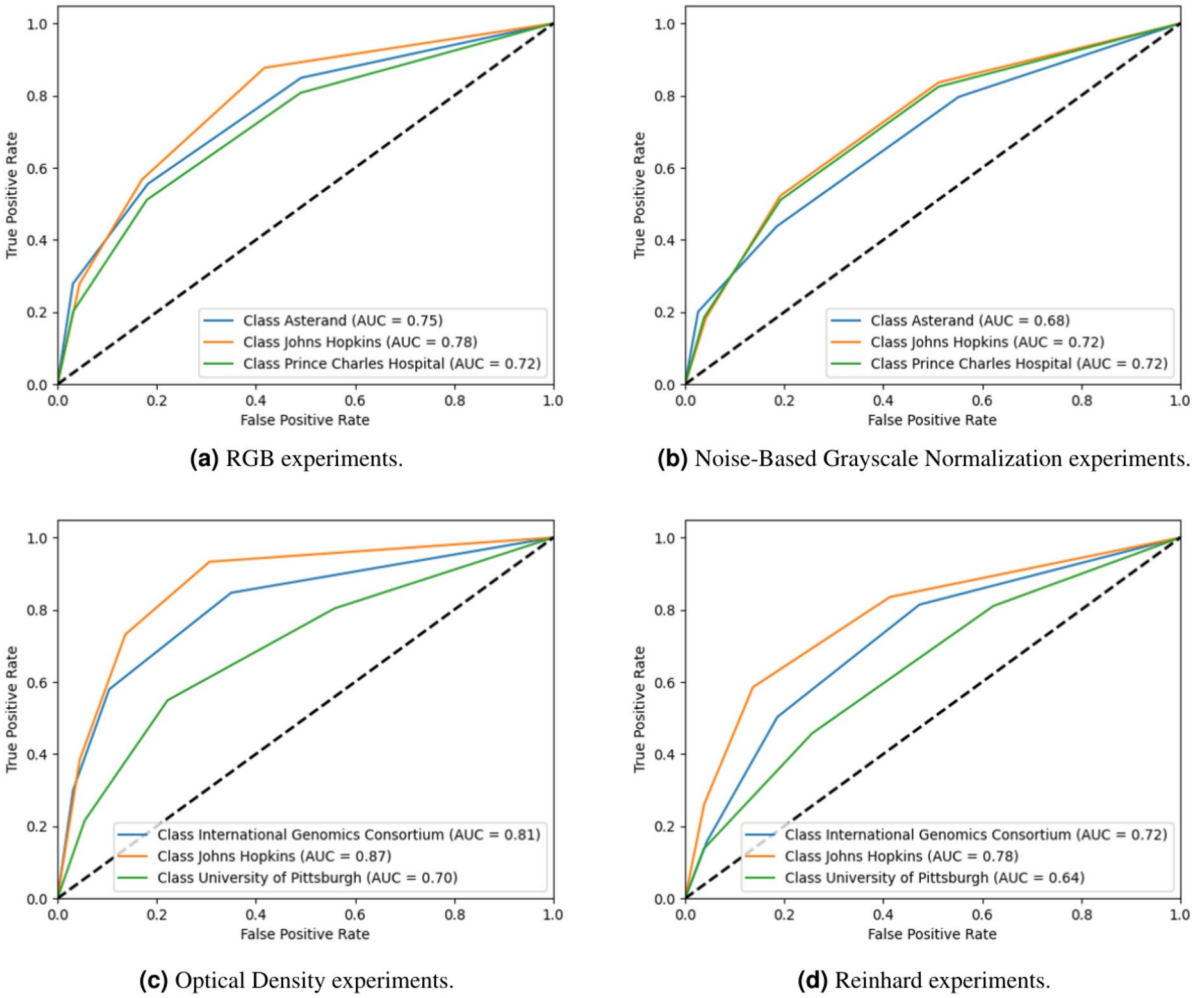
- **Cancer Classification:** For the RGB dataset, as shown in Fig. 5, the AUROC is 0.82 for both classes, indicating a strong discrimination capability for cancer detection. On the other hand, for the dataset after applying the NBGN method, the AUROC slightly decreases to 0.80, suggesting that the removal of color information has a minor impact on the model's ability to detect cancer while still maintaining good performance. For the OD

transformation method, the AUROC decreases to 0.73, indicating a notable drop in the model's discrimination capability, potentially due to the method's less effective handling of color variability and its impact on the cancer-related features. Similarly, for the Reinhard transformation method, the AUROC is 0.75, demonstrating a modest performance compared to OD, likely due to better normalization of color stains while preserving relevant cancer-specific information.

- **Data Center Classification:** For the RGB dataset, the AUROC values, as depicted in Fig. 6, for the three data centers are 0.75, 0.78, and 0.72, demonstrating moderate discrimination capability. These values suggest that color patterns, likely introduced by data center-specific staining procedures, play a role in embedding center-specific information. For the normalized dataset via NBGN, the AUROC values drop to 0.68, 0.72, and 0.72, showing that the removal of color significantly affects data center identification. This drop highlights the reliance of the model on color-related patterns for distinguishing between data centers. Regarding OD transformation effects on eliminating data center signatures, we obtained 0.81, 0.87 and 0.70 values, and for Reinhard transformation, these values are 0.72, 0.78 and 0.64. Overall, these results demonstrate the varying effectiveness of different stain normalization methods in mitigating center-specific biases. While the NBGN method in overall achieves more reduction compared to two other normalization methods.

The ROC analysis supports our hypothesis that color information introduces bias, particularly in data center classification, where center-specific staining patterns are utilized by the model. The more pronounced decrease in AUROC for data center identification compared to cancer classification further reinforces this finding. By removing color through transformations, the model's reliance on biased features is reduced.

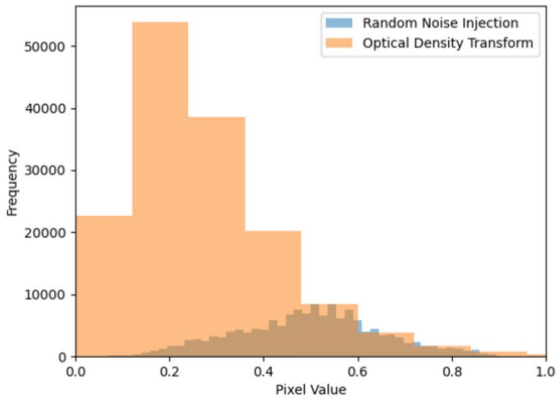
Data distribution analysis across transformation methods To analyze the impact of different stain normalization methods on the data distribution, we compared the histograms of pixel values for OD, Reinhard, and the NBGN transformations. The histograms provide insights into how each method modifies the data and redistributes pixel intensities.



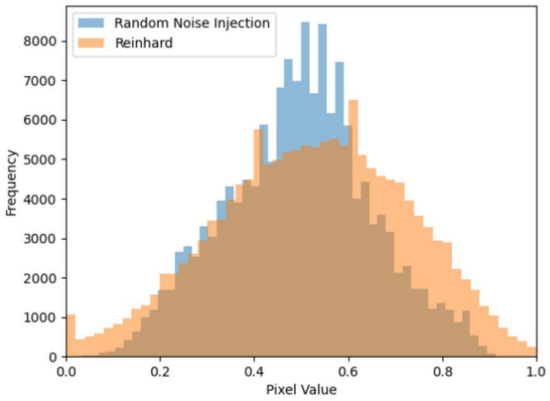
(a) RGB experiments. **(b)** Noise-Based Grayscale Normalization experiments.

(c) Optical Density experiments. **(d)** Reinhard experiments.

Fig. 6. ROC curves for multi-class data center classification under different experimental conditions: RGB (**a**), Noise-Based Grayscale Normalization (**b**), Optical Density (**c**), and Reinhard (**d**). The curves illustrate the model's performance in distinguishing data centers for each experiment.



(a) Optical Density Transformation vs. Noise-Based Grayscale Normalization



(b) Reinhard Transformation vs. Noise-Based Grayscale Normalization

Fig. 7. Histograms comparing the pixel value distributions of different transformation methods: (**a**) Optical Density transformation versus Noise-Based Grayscale Normalization and (**b**) Reinhard transformation versus Noise-Based Grayscale Normalization. The Optical Density transformation exhibits a compressed distribution with most pixel values concentrated in the lower range, while the Reinhard transformation achieves a Gaussian-like distribution. The Noise-Based Grayscale Normalization method produces a broader, more uniform distribution, reflecting its better ability to maintain target features while neutralizing center-specific patterns.

The histogram for the OD transformation, shown in Fig. 7a, reveals a strongly uneven distribution, with most pixel values concentrated in the lower intensity range (around 0.2). This indicates that the OD transformation tends to compress pixel intensities into a narrower range, potentially losing some fine-grained detail. In contrast, the NBGN method demonstrates a broader distribution, with pixel values more evenly spread across the intensity range, reflecting its ability to maintain variability by injecting noise. The difference in distributions suggests that the OD transformation may over-normalize the data, which could explain its limited effectiveness in mitigating data center-specific biases.

The histogram for the Reinhard transformation, shown in Fig. 7b, depicts a Gaussian-like distribution of pixel values centered around 0.5. The overlap between the two histograms indicates that Reinhard normalization effectively retains some variability in pixel intensities, making it a more balanced approach compared to the OD transformation. The smoother distribution achieved by the Reinhard transformation aligns with its performance in reducing center-specific biases while preserving key information. However, the NBGN method exhibits a more uniform distribution, which suggests it might be better at neutralizing specific patterns.

Discussion

The paper systematically investigates bias in deep learning models applied to histopathology image analysis, particularly focusing on the TCGA dataset. It identifies and examines the sources and impacts of bias, utilizing KimiaNet and EfficientNet for feature extraction and analysis. Through a series of tests and analyses, including data imbalance, preprocessing effects, and stain variation impacts, the study underscores the significance of addressing bias to enhance model reliability and generalizability.

The study began with an examination of KimiaNet's performance, where disparities in data balance and pre-processing steps were initially identified as potential sources of bias. The disparity analysis and the clustered heatmap visualization technique highlighted how data imbalance could affect the model's balanced accuracy value, suggesting that models might rely more on center-specific signatures rather than cancerous features themselves. Further investigations using the k-NN classifier revealed that a high proportion of nearest neighbors for correctly classified samples belonged to the same cancer type, indicating that the model might be using site-specific data-gathering protocols as shortcuts for cancer type discrimination.

To further investigate bias, the study switched to using EfficientNet on the raw TCGA dataset to explore additional sources of center-specific shortcuts. Initial experiments involved converting the dataset to grayscale after introducing controlled noise to assess the impact of staining variations on model performance. The results confirmed that site-specific staining patterns act as learned shortcuts, influencing model predictions. To further analyze the role of stain normalization in mitigating these biases, we applied two widely used normalization techniques-Reinhard and Optical Density transformations-alongside our Noise-Based Grayscale Normalization method. Each transformation was evaluated based on its effectiveness in reducing center-specific biases while preserving cancer classification accuracy. The results show that while all normalization techniques contributed to bias reduction, Reinhard and Noise-Based Grayscale Normalization methods were the most effective in balancing bias mitigation with diagnostic performance.

Despite these efforts, our research study is not without its limitations. One significant limitation of this study is the selection of a balanced dataset that includes only two cancer types. Achieving a balanced set with a broader range of cancer types is challenging due to the uneven distribution of samples in the original TCGA dataset.

Many cancer types are either restricted to a few data centers or lack sufficient samples, making it difficult to create a more representative balanced set. As a result, we were limited to using samples from only two cancer types, which may impact the generalizability of our findings. Furthermore, despite the efforts to identify the bias factors, the study acknowledges that site-specific bias could not be fully recognizable. This highlights a key limitation: even with interventions like noise injection and grayscale transformations, the model may still rely on site-specific shortcuts rather than learning purely from cancerous features.

However, the findings demonstrate the intricate nature of bias and the crucial need to conduct more research to enhance the reliability, fairness, and generalizability of AI models in healthcare diagnostics. In fact, there is a need for innovative strategies that can detect and correct biases during the model training phase without compromising the ability to detect clinically relevant features in histopathology images. However, in some cases, it is not possible to retrain the model in order to filter bias, like the features extracted by large and complex deep models like KimiaNet. In these cases, post-training methods, like feature selection, may be an efficient way to eliminate bias signatures from learned features. In addition, investigating different neural network architectures or hybrid models may provide insights into configurations that are less susceptible to biases inherent in current deep learning approaches. Another potential solution could be applying machine unlearning methods to remove irrelevant features. Machine unlearning, also referred to as data deletion or model unlearning, is an emerging area in machine learning that involves the ability to remove the influence of specific data points from a machine-learning model without completely retraining the model from scratch. Therefore, this area seems a more promising solution in addressing bias concerns in AI models used for histopathology and other applications.

Data availability

All the test cases were applied on the TCGA dataset publicity available at the following "https://portal.gdc.cancer.gov".

Code availability

The code packages for each case study are uploaded to the following GitHub repository for public access: GitHub Repository.

Received: 10 June 2024; Accepted: 4 February 2025
Published online: 02 April 2025

References

1. Gervasi, S. S. et al. The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Aff.* **41**, 212–218 (2022).
2. Caton, S. & Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* (2020).
3. Verma, S. & Rubin, J. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1–7 (2018).
4. Tian, H., Zhu, T., Liu, W. & Zhou, W. Image fairness in deep learning: Problems, models, and challenges. *Neural Comput. Appl.* **34**, 12875–12893 (2022).
5. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
7. Pagano, T. P. et al. Bias and unfairness in machine learning models: A systematic literature review. arXiv preprint arXiv:2202.08176 (2022).
8. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
9. Soneson, C., Gerster, S. & Delorenzi, M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS ONE* **9**, e100335 (2014).
10. Hägele, M. et al. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci. Rep.* **10**, 6423 (2020).
11. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).
12. Brown, A. et al. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat. Commun.* **14**, 4314 (2023).
13. Roach, M. *et al.* Prostate cancer risk in African American men evaluated via digital histopathology multi-modal deep learning models developed on NRG oncology phase III clinical trials (2022).
14. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
15. Vaidya, A. et al. Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.* **30**, 1174–1190 (2024).
16. Dhont, J., Wolfs, C. & Verhaegen, F. Automatic coronavirus disease 2019 diagnosis based on chest radiography and deep learning-success story or dataset bias?. *Med. Phys.* **49**, 978–987 (2022).
17. Arias-Garzón, D., Tabares-Soto, R., Bernal-Salcedo, J. & Ruz, G. A. Biases associated with database structure for covid-19 detection in x-ray images. *Sci. Rep.* **13**, 3477 (2023).
18. Correa, R. et al. A systematic review of 'fair' AI model development for image classification and prediction. *J. Med. Biol. Eng.* **42**, 816–827 (2022).
19. Zhou, Y. *et al.* Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. arXiv preprint arXiv:2111.11665 (2021).
20. Rotemberg, V. et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **8**, 34 (2021).
21. Correa, R. *et al.* Two-step adversarial debiasing with partial learning–medical image case-studies. arXiv preprint arXiv:2111.08711 (2021).
22. Dehkharghanian, T. et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn. Pathol.* **18**, 1–12 (2023).
23. Maleki, F. et al. Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *Radiol. Artif. Intell.* **5**, e220028 (2022).
24. Nauta, M., Walsh, R., Dubowski, A. & Seifert, C. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics* **12**, 40 (2021).

25. Mazaheri, P., Bidgoli, A. A., Rahnamayan, S. & Tizhoosh, H. R. Ranking loss and sequestering learning for reducing image search bias in histopathology. *Appl. Soft Comput.* **142**, 110346 (2023).
26. Asilian Bidgoli, A., Rahnamayan, S., Dehkharghanian, T., Grami, A. & Tizhoosh, H. R. Bias reduction in representation of histopathology images using deep feature selection. *Sci. Rep.* **12**, 19994 (2022).
27. Kheiri, F., Bidgoli, A. A., Makrehchi, M. & Rahnamayan, S. Feature selection-driven bias deduction in histopathology images: Tackling site-specific influences. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, 1–8 (IEEE, 2024).
28. Riasatian, A. et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021).
29. Gutman, D. A. et al. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J. Am. Med. Inform. Assoc.* **20**, 1091–1098 (2013).
30. Kalra, S. et al. Yottixel-an image search engine for large archives of histopathology whole slide images. *Med. Image Anal.* **65**, 101757 (2020).
31. Koonce, B. & Koonce, B. Efficientnet. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization* 109–123 (2021).
32. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69**, 066138 (2004).
33. Hoque, M. Z., Keskinarkaus, A., Nyberg, P. & Seppänen, T. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Inf. Fusion* 101997 (2023).
34. Yengec-Tasdemir, S. B., Aydin, Z., Akay, E., Dogan, S. & Yilmaz, B. Improved classification of colorectal polyps on histopathological images with ensemble learning and stain normalization. *Comput. Methods Programs Biomed.* **232**, 107441 (2023).
35. Ruifrok, A. C. et al. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
36. Roy, S., Kumarjain, A., Lal, S. & Kini, J. A study about color normalization methods for histopathology images. *Micron* **114**, 42–61 (2018).

Author contributions

F.K. is the corresponding author. S.R. conceived the study. M.M. and S.R. supervised the test case design. S.R. and A.A.B. reviewed and edited the manuscript. F.K. made contributions to the design and implementation of test cases, results analysis, and manuscript drafting. All authors read and approved the final manuscript.

Declarations

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/s41598-025-89210-x.

Correspondence and requests for materials should be addressed to F.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.