Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250302

Bias Detection in Histology Images Using Explainable AI and Image Darkness Assessment

Inna SKARGA-BANDUROVA[a,1], Golshid SHARIFNIA[a], and Tetiana BILOBORODOVA[b]
[a] *Oxford Brookes University*
[b] *G.E. Pukhov Institute for Modelling in Energy Engineering*
ORCiD ID: Inna Skarga-Bandurova https://orcid.org/0000-0003-3458-8730, Golshid Sharifnia https://orcid.org/0000-0002-0993-9458, Tetiana Biloborodova https://orcid.org/0000-0001-7561-7484

Abstract. The study underscores the importance of addressing biases in medical AI models to improve fairness, generalizability, and clinical utility. In this paper, we present a novel framework that combines Explainable AI (XAI) with image darkness assessment to detect and mitigate bias in cervical histology image classification. Four deep learning architectures were employed—AlexNet, ResNet-50, EfficientNet-B0, and DenseNet-121—with EfficientNet-B0 demonstrating the highest accuracy post-mitigation. Grad-CAM and saliency maps were used to identify biases in the models' predictions. After applying brightness normalisation and synthetic data augmentation, the models shifted focus toward clinically relevant features, improving both accuracy and fairness. Statistical analysis using ANOVA confirmed a reduction in the influence of image darkness on model predictions after mitigation, as evidenced by a decrease in the F-statistic from 120.79 to 14.05, indicating improved alignment of the models with clinically relevant features.

Keywords. Explainable AI (XAI), histopathology, bias detection, image darkness

1. Introduction

Artificial intelligence (AI) has become integral to modern healthcare, particularly in diagnostic imaging, where it assists clinicians in analysing large volumes of data quickly and accurately [1]. However, despite these advancements, the potential for bias in AI models remains a significant concern, particularly when models are trained on biased datasets. Bias in medical datasets can arise from several factors, such as variations in data acquisition techniques, image quality, or under-representation of certain classes within the dataset [2,3]. These biases can manifest in AI models, leading to poor generalisation, skewed predictions, and, ultimately, suboptimal patient outcomes. For instance, in the case of histology images, differences in staining intensity, tissue preparation techniques, and image brightness or darkness across institutions can introduce biases that cause the model to misclassify them [4]. Thus, understanding the model's behaviour could reveal that it is paying more attention to darker or brighter

[1] Corresponding Author: Inna Skarga-Bandurova; E-mail: iskarga-bandurova@brookes.ac.uk

regions that are artefacts rather than diagnostic features. Explainable AI (XAI) offers a potential solution to this problem by providing insights into the model's decision-making process. While XAI methods have traditionally been used to make AI predictions more interpretable, combining them with darkness analysis could highlight how the model's predictions are influenced by poor image quality to improve robustness. This idea is justified by broader research on XAI and image quality biases [5] along with the colour-based computations [6].

This study presents a novel framework that combines image darkness assessment with XAI to detect and mitigate bias in cervical histology image classification. By assessing image quality variations and identifying biased patterns in the model's predictions, this approach ensures that the model focuses on clinically relevant features rather than irrelevant artefacts. Through this approach, we aim to enhance the fairness, transparency, and generalizability of AI models in medical imaging, ensuring they deliver reliable and unbiased outcomes in clinical practice. The key contribution of this work lies in the use of XAI not only as a post-hoc interpretability tool but also as a proactive method for identifying and addressing dataset biases.

2. Methods

The dataset [7] used in this study consists of cervical histology images labelled into four categories: norm, CIN1, CIN2, and CIN3. These images were captured from different institutions, introducing inherent variability in staining intensity and image quality. To account for this, the dataset was pre-processed and divided into a training set (80%) and a testing set (20%). Data augmentation techniques, including random rotations, colour jittering, and horizontal flipping, were applied to the training set to simulate variations in staining and improve the model's ability to generalise. The images were also resized, centre-cropped, and normalised to ensure consistency during training.

Four deep learning architectures were selected for this study: AlexNet [8], ResNet-50 [9], EfficientNet-B0 [10] and DenseNet-121 [11]. The first two architectures were chosen due to their proven success in image classification, with ResNet-50 offering deeper feature extraction capabilities through its residual connections. EfficientNet-B0 was chosen as the primary model due to its compound scaling approach, which balances depth, width, and resolution, allowing the model to achieve higher accuracy with fewer parameters. DenseNet-121 was selected because of its dense connectivity between layers, which promotes better gradient flow and prevents overfitting, making it ideal for handling imbalanced datasets. All models were initialised using pre-trained weights from ImageNet, enabling transfer learning. Fine-tuning was applied to the final layers to adapt the models to the histology classification task, modifying the output layers to align with the four classes. AlexNet and ResNet-50 models were trained for 24 epochs using a batch size of 32, with a starting learning rate of $1e^{-4}$. A learning rate decay was applied every 10 epochs to avoid overfitting. The final fully connected layers of EfficientNet were replaced with new layers specific to the CIN classification task. Early stopping was applied to prevent overfitting during training. In DenseNet, the last few layers were fine-tuned, and dropout layers were added to reduce overfitting by randomly dropping neurons during training. To understand how the models made decisions and to identify potential biases, two explainability techniques were utilised: Grad-CAM and saliency maps. Grad-CAM was employed to generate heatmaps that visualise the regions of the image the model deems most important for predictions. It allows for visually inspect

whether the model focuses on pathologically relevant features or irrelevant artefacts. Saliency maps were used to provide a more granular view of the model's decision-making process to identify biases at the pixel level.

A key component of this study is the assessment of image darkness as a potential source of bias. Since staining intensity varies between institutions, we hypothesised that the models might rely on image darkness to differentiate between CIN grades rather than clinically relevant features. Each histology image was first converted to grayscale, and the average pixel intensity was calculated to determine the darkness value. This value was then used to assess the relationship between image darkness and the predicted CIN class. The combination of XAI and image darkness analysis was used to detect whether the model's decisions were influenced by non-pathological factors, such as overly dark regions caused by staining inconsistencies. Once biases were detected, additional data augmentation techniques were employed to mitigate them. Images were normalised to a consistent brightness range to reduce the influence of staining intensity. Synthetic images with varied staining intensities were generated to teach the model to focus on relevant tissue features independent of darkness or brightness artefacts.

To assess the relationship between image darkness and the predicted CIN class, a one-way analysis (ANOVA) was used, where the predicted class served as the dependent variable, while image darkness (average pixel intensity) was the independent variable. This allowed us to determine whether variations in image darkness affected the model's predictions, thereby indicating potential bias. A significant association suggests the model might be relying on artifacts instead of focusing on clinically relevant structures.

The performance of the XAI techniques was qualitatively evaluated by pathologists, who assessed whether the heatmaps and pixel-level explanations highlighted clinically significant areas of the histology images (e.g., nuclear abnormalities).

3. Results

3.1. Baseline Model Performance

All four models were trained and evaluated on the cervical histology dataset prior to any bias mitigation. EfficientNet-B0 demonstrated the highest overall accuracy at 88%, surpassing the AlexNet architecture, which achieved 80% accuracy, also being less effective at classifying CIN1. ResNet-50 and DenseNet-121 performed similarly, achieving 86% and 87% accuracy, respectively, with high precision and recall, particularly for CIN2 and CIN3 classes, indicating their ability to accurately distinguish between higher-grade lesions.

3.2. Explainability Insights

Visualisations produced by Grad-CAM (Fig. 1 (d)) revealed that all four models tended to focus on darker regions in certain images, particularly in the CIN1 and CIN2 classes. This suggests that the models may have relied on variations in staining intensity rather than relevant tissue structures. Saliency maps further confirmed this finding, showing that dark regions were disproportionately influencing the model's decisions. The models focused on non-pathological features (artefacts caused by inconsistent staining) in some instances rather than morphological features indicative of CIN stages. Fig. 1 (a), (b) show that the model identified cell nuclei as the elements most responsible for the severity of

the degree, while (c), and (d) show that the model also identifies tissue artefacts and areas of more intense colour as highly relevant.

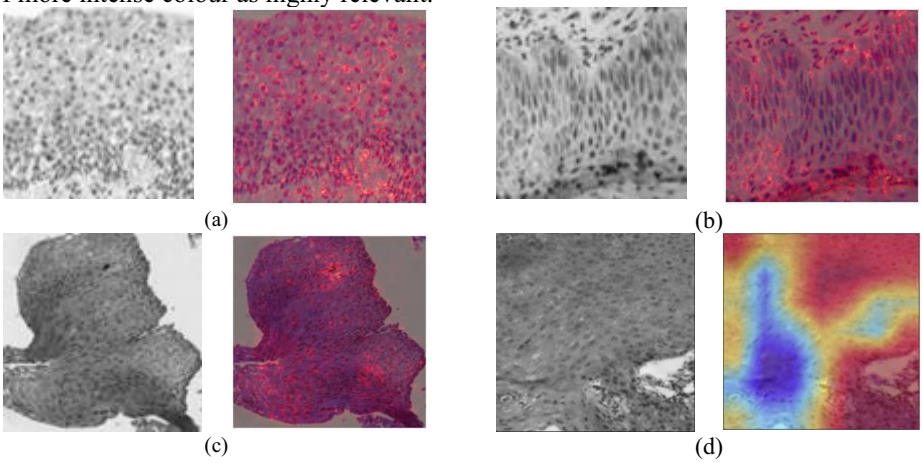

Figure 1. CIN images converted to grayscale and corresponding saliency maps (a)-(c) and Grad-CAM (d).

3.3. Image Darkness and Bias Detection

The grayscale conversion and subsequent darkness value calculation revealed significant variability in the average pixel intensity across the dataset. This variability was associated with the model predictions, particularly for the CIN1 and CIN2 categories, which exhibited the highest variance in darkness. The ANOVA results revealed a statistically significant effect of image darkness on CIN model predictions (F-statistic=120.79, p=2.475e-40). This indicates that the mean image darkness differs significantly across CIN classes, suggesting that image darkness is influencing the model's predictions. Such reliance on image darkness, rather than solely on clinically relevant histopathological features, confirms the presence of bias due to variations in staining intensity.

3.4. Bias Mitigation Results

After detecting the biases, several bias mitigation techniques were applied, including brightness normalisation and data augmentation (introducing synthetic staining variations). The models were then retrained on the updated dataset, and their performance was reassessed. EfficientNet-B0 achieved the highest accuracy post-bias mitigation, with an increase from 88% to 91%. The F1-score for the CIN1 and CIN2 categories showed the most significant improvement, confirming that the model was no longer overly reliant on staining intensity variations. ResNet-50 and DenseNet-121 also showed improvements in both accuracy and recall, with ResNet-50 increasing from 86% to 88% and DenseNet-121 increasing to 88%. These results demonstrate the robustness of these models after mitigating the impact of image quality biases.

3.5. Explainability Insights and Reduction in Image Darkness Bias

The Grad-CAM heatmaps produced after applying brightness normalisation showed a significant shift in the model’s focus. The models concentrated more on clinically

relevant regions, such as cell nuclei and epithelial structures, rather than on dark staining artefacts. Similarly, the saliency maps revealed a more even distribution of pixel importance, with reduced reliance on darkened regions of the images. A second ANOVA test confirmed that the association between image darkness and the predicted CIN class had significantly diminished after bias mitigation (F-statistic=14.05, p=0.031). This demonstrates that the models relied less on staining artefacts to make predictions and instead focused on histopathology-relevant features. Analysis of the variance in model predictions before and after image normalisation showed that the bias mitigation strategies significantly reduced the variance for the CIN1 and CIN2 classes, confirming a reduction in model over-reliance on non-pathological features.

4. Discussion and Conclusions

The study introduced a novel framework for detecting and mitigating bias in cervical histology image classification using a combination of XAI with image darkness assessment. The goal was to ensure that AI models focus on clinically relevant features, such as nuclear abnormalities and epithelial structures, rather than irrelevant artefacts introduced by staining variations or image darkness. The results demonstrated that the models were influenced by non-pathological features such as staining intensity, particularly in the CIN1 and CIN2 categories. After applying data augmentation and brightness normalisation, the models' focus shifted towards more relevant anatomical features. The ANOVA tests further confirmed a reduction in the influence of image darkness on model predictions post-mitigation, with the F-statistic decreasing from 120.79 to 14.05. This indicates that the models were less reliant on staining artefacts and better aligned with clinically relevant features.

References

[1] Alowais SA, Alghamdi SS, Alsuhebany N et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 23, 689 (2023). doi: 10.1186/s12909-023-04698-z.
[2] Vargas-Cardona HD, et al. Artificial intelligence for cervical cancer screening: Scoping review, 2009-2022. Int J Gynaecol Obstet. 2024 May;165(2):566-578. doi: 10.1002/ijgo.15179.
[3] Drukker K, Chen W, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham). 2023 Nov;10(6):061104. doi: 10.1117/1.JMI.10.6.061104.
[4] Evans H, Snead D. Why do errors arise in artificial intelligence diagnostic tools in histopathology and how can we minimize them? Histopathology, 2024; 84(2), pp.279-287. doi: 10.1111/his.15071.
[5] Shahbazi N, Lin Y, Asudeh A, Jagadish HV. Representation Bias in Data: A survey on identification and resolution techniques. ACM Comput. Surv. 2023; 55(13). doi: 10.1145/3588433.
[6] Timchenko V, et al. Effectiveness evaluations of optical color fuzzy computing. In: Research tendencies and prospect domains for AI development and implementation, 2024; 129-150. River Publishers.
[7] Brosnan B, et al. Cervical Intraepithelial Neoplasia Grading from Prepared Digital Histology Images. Stud Health Technol Inform. 2023 Jun 29; 305:402-405. doi: 10.3233/SHTI230516. PMID: 37387050.
[8] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM, 2017; 60, 6 (June 2017), 84–90. https://doi.org/10.1145/3065386.
[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.
[10] Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv, 2019; abs/1905.11946.
[11] Huang G, Liu Z, Van Der Maaten L, Weinberger K. Densely Connected Convolutional Networks. ArXiv, 2016; abs/1608.06993.