




OPEN

Improving quality control of whole slide images by explicit artifact augmentation

Artur Jurgas^{1,2}, Marek Wodzinski^{1,2}, Marina D'Amato³, Jeroen van der Laak³,
Manfredo Atzori^{2,4} & Henning Müller^{2,5}

The problem of artifacts in whole slide image acquisition, prevalent in both clinical workflows and research-oriented settings, necessitates human intervention and re-scanning. Overcoming this challenge requires developing quality control algorithms, that are hindered by the limited availability of relevant annotated data in histopathology. The manual annotation of ground-truth for artifact detection methods is expensive and time-consuming. This work addresses the issue by proposing a method dedicated to augmenting whole slide images with artifacts. The tool seamlessly generates and blends artifacts from an external library to a given histopathology dataset. The augmented datasets are then utilized to train artifact classification methods. The evaluation shows their usefulness in classification of the artifacts, where they show an improvement from 0.10 to 0.01 AUROC depending on the artifact type. The framework, model, weights, and ground-truth annotations are freely released to facilitate open science and reproducible research.

Keywords Deep learning, Quality assurance, Computed histopathology

With advancements in medical imaging techniques, histopathology remains an irreplaceable cornerstone, offering vital information for disease diagnosis, treatment decisions, and prognosis evaluation. This discipline forms the foundations of various medical specialties, including pathology, oncology, dermatology, and more. With the continued rise of computed histopathology, it carries an increased importance of quality control (QC) systems¹.

Histopathology not only offers a microscopic view of tissue structure but thanks to dyes that stain different structures, can also impart invaluable information about the tissue's composition. This unique combination of ultra-high resolution and color detail distinguishes histopathology from other imaging modalities. However, due to its intricate nature, histopathology workflows are susceptible to various artifacts, like shown in Fig. 1. The artifacts are unwanted structures or inconsistencies that obstruct the clear visualization of tissue structures².

Artifacts in histopathological images come from various sources, such as flaws in slide preparation, imperfect aperture settings during imaging. They can also come from improperly carrying out staining procedures, especially when it comes to immunohistochemistry dyes^{3,4}. Inconsistent staining, caused by variations in reagent concentrations, incubation times, or temperature fluctuations, can result in uneven coloration, obscuring critical cellular details^{5,6}. Even the most subtle discrepancies in these processes can give rise to artifacts within Whole Slide Images (WSIs), which can potentially compromise the accuracy of medical diagnoses.

Given the significance of histopathology in contemporary medical workflows, it is imperative to ensure that the acquired images are of the highest quality possible. The need for rigorous quality control (QC) mechanisms in histopathology is evident, as the precision of medical diagnoses, treatment planning, and patient outcomes hinge on the reliability of these images. This includes early detection, localization, and classification of unwanted structures on the image, especially if they obstruct the view of the tissue being examined. Structures like that need to be recognized and acknowledged, ideally during the acquisition process as early as possible. The problem is further complicated by the size of the image, as well as the high heterogeneity of the artifacts themselves⁷. This paper focuses on the development and application of a fully end-to-end pipeline for training deep learning-based quality control systems that addresses the challenge of artifact detection in histopathological WSIs, enhancing the reliability and accuracy of medical interpretations.

¹AGH University of Krakow, Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, 30059 Krakow, Poland. ²University of Applied Sciences Western Switzerland (HES-SO), Institute of Informatics, 3960 Sierre, Switzerland. ³Radboud University Medical Center, Nijmegen, The Netherlands. ⁴Department of Neuroscience, University of Padova, Padua, Italy. ⁵Medical Faculty, University of Geneva, Geneva, Switzerland. ✉email: arjurgas@agh.edu.pl

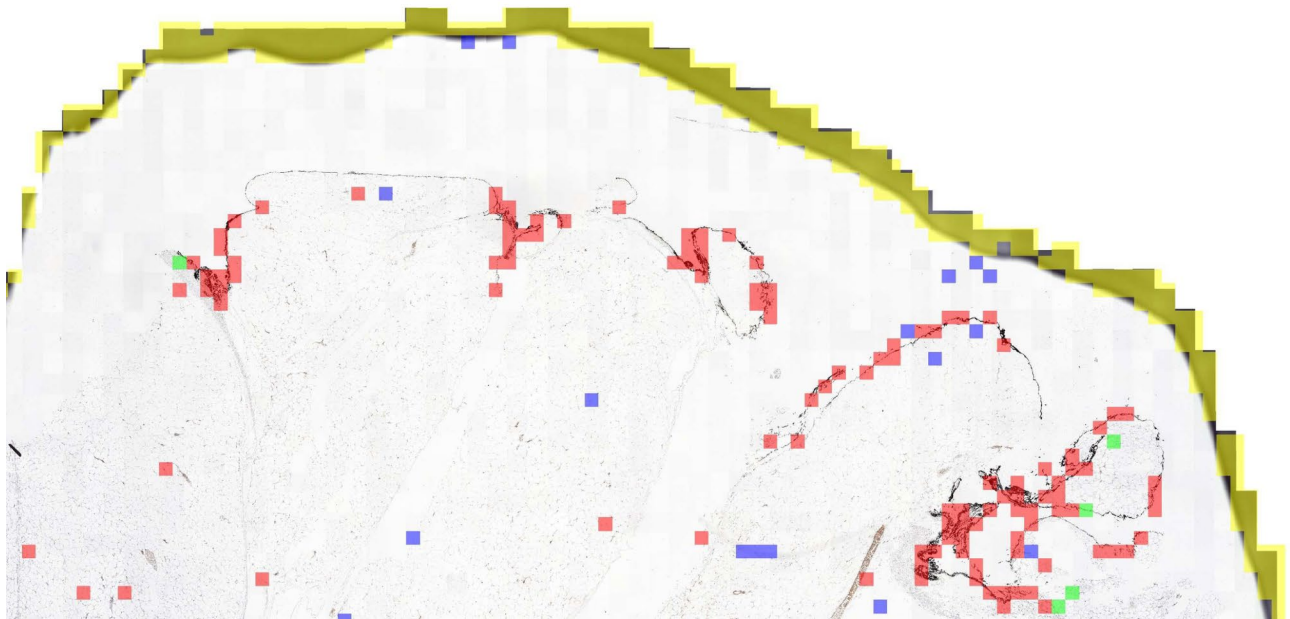


Figure 1. Aggregation of a whole slide image patch-level classification. Color meaning: (yellow) marker, (red) ink, (blue) dust, (green) tissue folding.

Recent developments in histopathology QC systems have shown promising strides towards automated assessment. Notably, the semi-automatic HistoQC software^{8,9} facilitates the automatic evaluation of WSIs for quality assessment (e.g., pen markings, air bubbles, blur), yet several challenges still persist. A primary concern arises from the high heterogeneity observed in clinical datasets, necessitating advanced artifact detection algorithms for each artifact type. This leads to significant time demands in both development and inference phases.

Most algorithms' effectiveness is prominent in the context of Hematoxylin and Eosin (H &E) staining, but can exhibit limitations when applied to Immunohistochemistry (IHC) staining. This is due to those dyes' unique challenges and variations, as presented in Fig. 2. Moreover, most software is parametrized, requires examples of artifacts, and training classification models for each inference. As per the official documentation of HistoQC¹⁰, it is also recommended to work with images at 1.25× magnification; otherwise the processing can become infeasible. HistoQC is a great achievement for an end-to-end WSI analysis platform, where our method could potentially improve the artifact detection part of the pipeline.

Other authors have explored learning-based QC systems^{11,12}, but their capabilities are largely confined to detecting blur and out-of-focus artifacts, which are relatively easier to generate synthetically, where they convolve parts of an image with a Gaussian function. According to¹³ and¹⁴, specific models have been trained to detect tissue folds and ink artifacts, respectively. These models show good performance, but are evaluated only on a limited dataset consisting of 79 and 89 H &E stained WSIs. The authors have not made the training data publicly available. In a supervised, learning-based approach¹⁵, a proprietary labeled dataset was employed for training, albeit with considerable cost and limited generalizability owing to extensive manual annotations. A similar work by¹⁶ leveraged deep residual networks but relied on a modest manually annotated dataset. The critical role of QC in learning-based histopathology methods is underscored in^{17,18}. Those works include the generation of artifacts

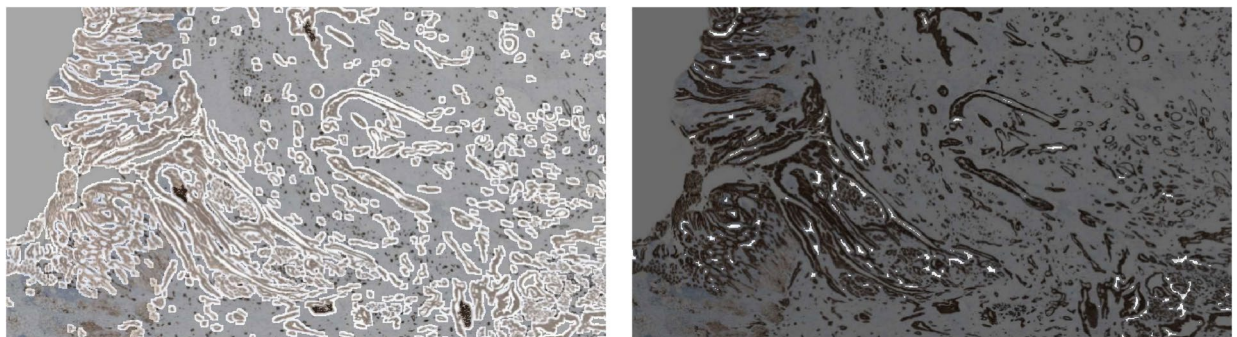


Figure 2. Example of failure of a quality control algorithm in identifying artifacts within an IHC image: (left) coverslip edge, (right) small objects (e.g., dust).

for stress-testing already trained deep learning models for evaluation. Synthetic artifacts are a prevalent focus there, with fewer non-synthetic artifacts addressed.

To this end, we introduce a framework aimed at generating representative and diverse datasets that mimic real-world scenarios. This is achieved by leveraging a limited number of annotations. We progressively apply these annotations to new, potentially unannotated data, thus constructing high-quality histopathology datasets for training segmentation and detection networks. This process involves extracting the artifact from the original image and seamlessly integrating it with images from other datasets. Consequently, the networks are trained to recognize this type of artifact across various backgrounds and tissue types. This significantly enhances the model's ability to generalize in comparison to basic train-time augmentation techniques based on geometry and intensity.

Contribution

This work presents a novel data augmentation method that extends high-resolution datasets with seamlessly blended real artifacts, fostering more realistic histopathological analysis challenges. We propose a methodology that involves the extraction of a representative sample of annotated artifacts, utilizing them to generate synthetic, realistic datasets for artifact detection and classification. By reducing the reliance on extensive professional annotations, this approach minimizes costs and enhances results. It also enables easy fine-tuning to accommodate specific institutional settings, allowing researchers to personalize and optimize artifact detection models. Additionally, the study demonstrates how this data generation method enhances the generalization of automatic, learning-based QC methods, resulting in improved performance and robustness.

Results

Datasets

The datasets employed in this study offer a diverse and comprehensive representation of histopathological artifacts. The ACROBAT challenge dataset¹⁹ comprises digitalized Whole Slide Images (WSIs) from FFPE surgical resection specimens of female primary breast cancer patients. Captured at 40X magnification (0.23 μm per pixel), these images, obtained using Hamamatsu Nanozoomer XR or Nanozoomer S360 scanners, exhibit a rich variety of artifacts. The evaluation focused on the validation subset, consisting of 100 cases equally divided between H &E and IHC-stained images.

Similarly, the ANHIR challenge dataset²⁰ provides a wide-ranging collection encompassing various tissues and pathological conditions, including lesions, lung lobes, mammary glands, colon adenocarcinoma (COAD), mice kidney tissue, gastric mucosa, gastric adenocarcinoma tissue, breast tissue, and kidney tissue. This dataset incorporates diverse staining techniques, employing stains such as Clara cell 10 protein, proSPC, H &E, Ki-67, PECAM-1, HER-2/neu, ER, PR, cytokeratin, and podocin. Acquired from various microscopy setups and scanners like Zeiss, Leica, 3DHistec, and NanoZoomer, the ANHIR dataset exhibits high heterogeneity, encompassing magnifications ranging from 10 \times to 40 \times and pixel sizes from 0.174 $\mu\text{m}/\text{pixel}$ to 2.294 $\mu\text{m}/\text{pixel}$.

Additionally, the Radboud University dataset, provided for evaluation purposes, stands as the largest both in artifact count and resolution. This dataset features professionally annotated artifacts in WSIs stained with both H &E and IHC dyes. Exemplary artifacts are present in Fig. 3. Spanning various tissue types, including bone marrow, breast tissue, colon tissue, pancreas tissue, diffuse large B-cell lymphoma (DLBCL), and images from the CAMELYON dataset²¹, each tissue type is characterized by different staining types, contributing to the dataset's richness and complexity. The Radboud University dataset offers a substantial resource for evaluating and validating the proposed quality control and segmentation methodologies in diverse histopathological contexts.

Selected artifacts are as follows:

- (i) Air. Not strictly connected to the tissue. Due to the fact that only a portion of the air bubble is frequently visible in the image, it assumes an open shape, often deviating from a complete circle.
- (ii) Dust. Small particles or debris that can inadvertently appear on the slides during the preparation or scanning process. It appears both on the foreground and background of the WSI.
- (iii) Tissue. Folded or creased tissue sections that can result from various factors such as handling, processing, or mounting of the tissue slides.
- (iv) Ink. Irregularities in the distribution or application of ink or staining agents on tissue slides.
- (v) Marker. Annotations, such as crosses or other symbols, typically located near the corners or edges of the slide.
- (vi) Focus. This artifact occurs when the focal plane of the microscope is not precisely aligned with the tissue section being captured, resulting in blurred or out-of-focus areas.

We summarize acquired data in Table 1.

Experimental setup

The experimental setup utilized Nvidia Tesla A100 graphics cards with 400W TDP and 40 GB of memory on the PLGrid HPC cluster Athena for model training. In our experiments, we employed deep learning models trained on different datasets, denoted by shorthand notations. After annotating a limited set of training artifacts, we utilize our framework to integrate them with other images. This results in the augmented datasets. The test datasets were unmodified. Models trained exclusively on annotated data from the ACROBAT dataset are referenced as **ACR**, while models trained on an augmented version of the ACROBAT dataset are denoted as **ACR'**. Similarly, with **ANH**, **ANH'** for ANHIR and **RB**, **RB'** for Radboud. Additionally, we evaluated models trained on ACROBAT datasets on the ANHIR dataset's annotations to analyze generalizability. Those models are denoted as **ACR_{anh}**.

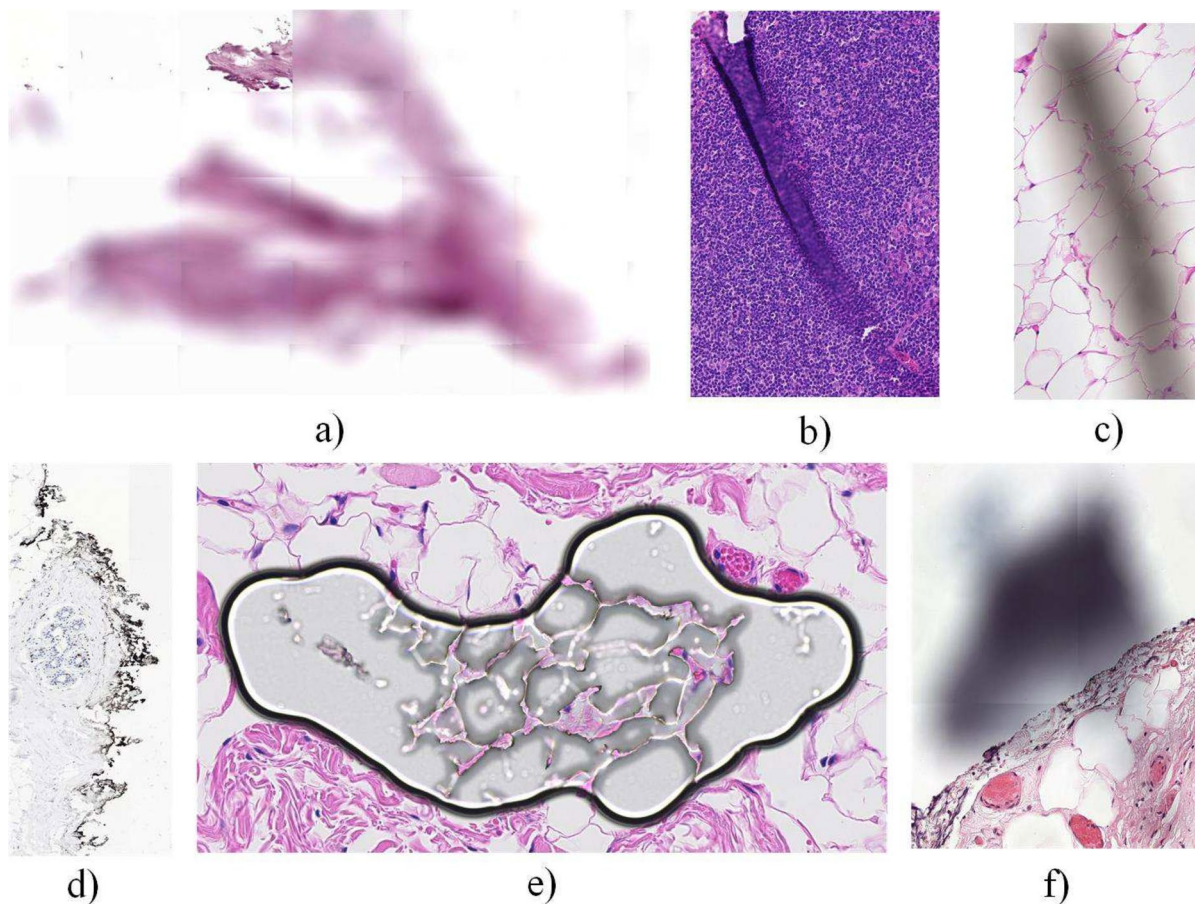


Figure 3. Examples of artifacts from the considered datasets: (a) focus, (b) tissue, (c) dust, (d) ink, (e) air, (f) marker.

Dataset	Source	Number of artifacts						
		Air	Dust	Tissue	Ink	Marker	Focus	All
ACROBAT	Self	67	137	89	87	97	14	491
ANHIR	Self	17	130	62	44	19	25	297
Radboud	Professional	149	398	1469	456	50	96	2618

Table 1. Datasets used in the study with their respective characteristics.

and, ACR'_{anh} respectively. When training the models on a dataset from the Radboud University, we present two approaches: (i) while having the full model set to trainable—**RB**, and (ii) only the last layers unfrozen—**RB_s**.

Classification

The classification study is presented through Receiver Operating Characteristic (ROC) curves accompanied by their corresponding Areas Under the Curve (AUC) scores (Table 2), offering a comprehensive evaluation of the models' performance. Figure 4 illustrates the promising initial validation results, with improvements evident when employing the augmented dataset, particularly in addressing previously weaker outcomes. Figure 5 details the loss on the validation dataset, highlighting the mitigation of overfitting issues with the augmented dataset during the training process. Subsequent testing on additional ACROBAT annotations reveals improvements for tissue and dust artifacts, alongside a performance decrease for ink artifacts and a slight drop in focus artifacts. Evaluation on ANHIR annotations demonstrates enhancements for air, tissue, dust, and focus artifacts, tempered by slight degradation in marker and ink artifacts.

In Fig. 6, the models undergo evaluation on a diverse set of WSIs. The augmented dataset yields improvements for air and dust artifacts, with more significant enhancements for dust, focus, and tissue types. However, a slight degradation is observed for ink and marker types. Improvement in performance on this dataset is the lowest overall. Further analysis in Table 3 reveals that the model does not generalize well to a new dataset. The lack of

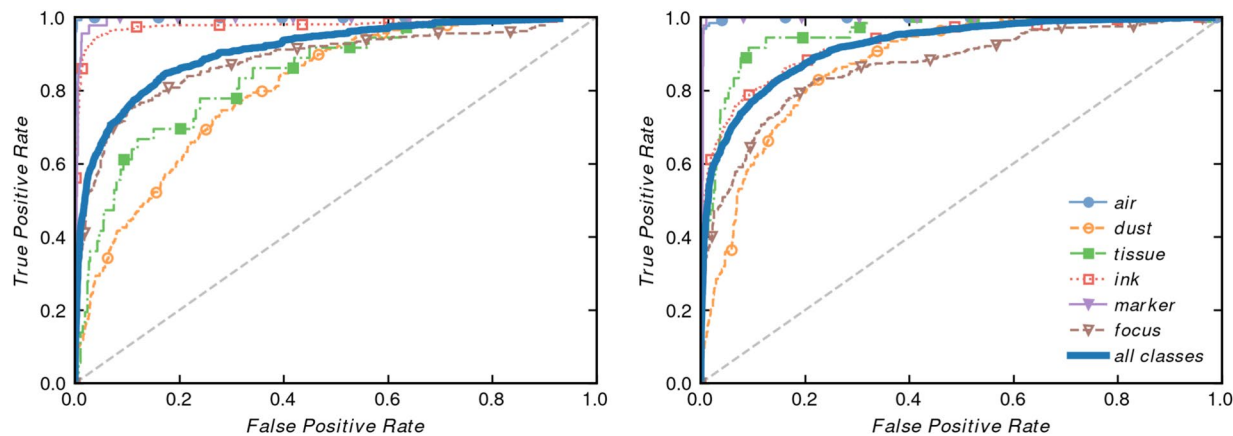


Figure 4. ROC curve for classification, evaluated on additional ACROBAT annotations unseen during training. (left) Model trained on ACR. (right) Model trained on augmented ACR'.

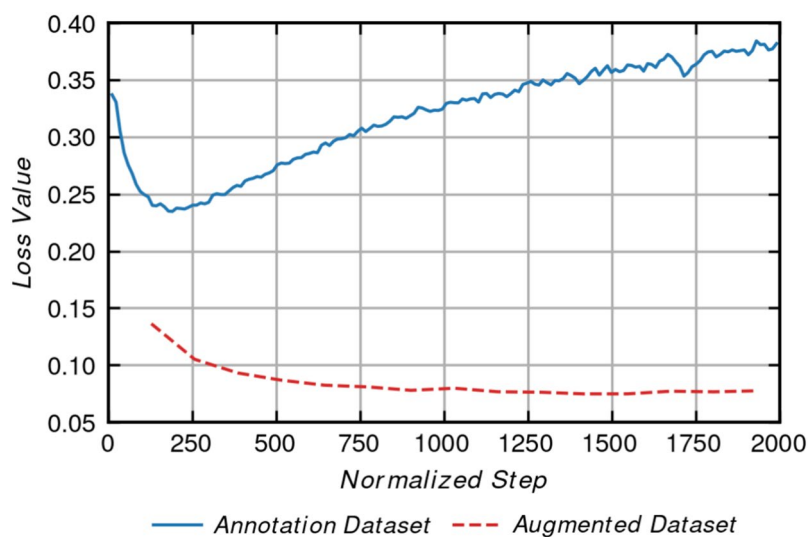


Figure 5. Chart of the validation loss during training on ACR and ACR' datasets.

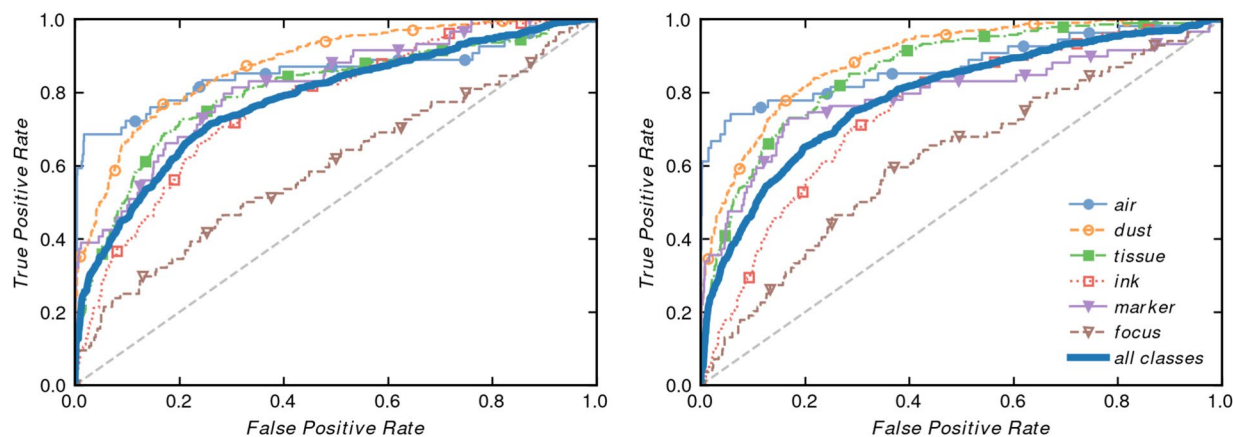


Figure 6. ROC curve for classification models, evaluated on ANHIR annotations unseen during training. (left) Model trained on ACR_{anh} . (right) Model trained on augmented ACR'_{anh} .

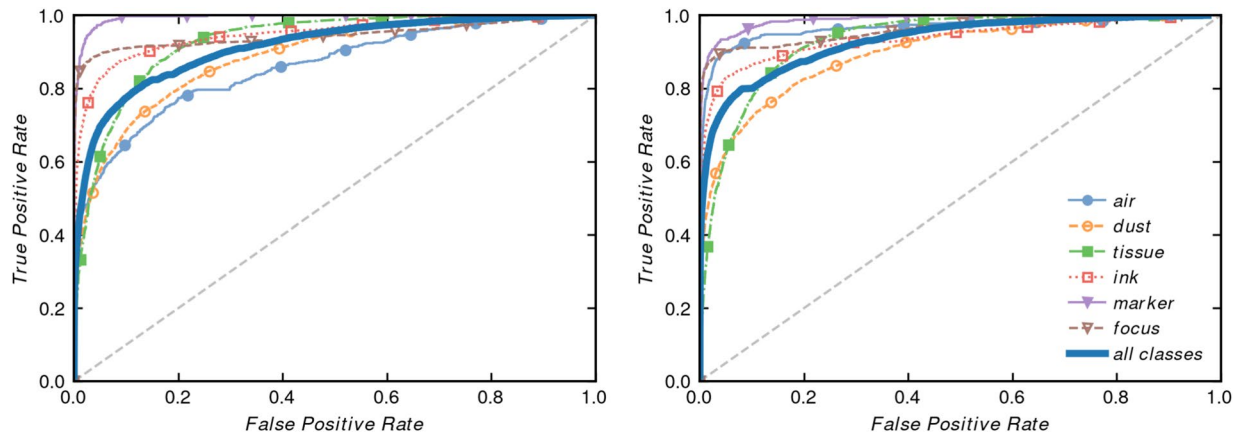


Figure 7. ROC curve for classification models, evaluated on Radboud University test annotations consisting of evenly sampled 70% of all dataset annotations. (left) Model trained on RB. (right) Model trained on augmented RB'.

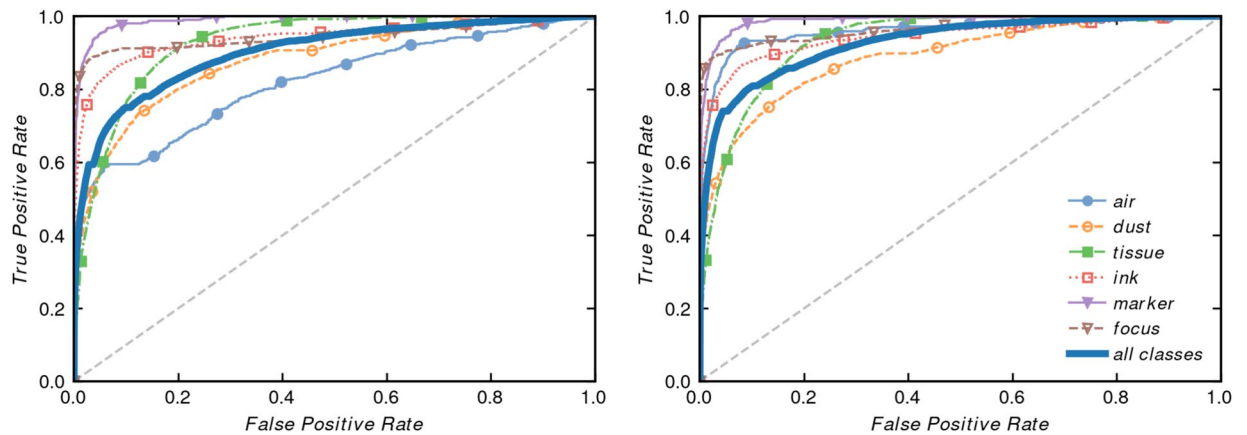


Figure 8. ROC curve for classification models, evaluated on Radboud University test annotations consisting of evenly sampled 70% of all dataset annotations. Model training was limited to only the last fully connected layer. (left) Model trained on RB_s. (right) Model trained on augmented RB'_s.

Artifact type	ACR	ACR'	ACR _{anh}	ACR' _{anh}	RB	RB'	RB _s	RB' _s
Air	1.000	0.999	0.858	0.872	0.858	0.965	0.820	0.961
Dust	0.807	0.876	0.876	0.892	0.887	0.898	0.880	0.889
Tissue	0.843	0.957	0.803	0.859	0.928	0.932	0.929	0.934
Ink	0.978	0.930	0.770	0.753	0.947	0.936	0.940	0.943
Marker	0.995	0.998	0.820	0.791	0.994	0.988	0.989	0.989
Focus	0.880	0.867	0.597	0.621	0.945	0.963	0.944	0.966
All classes	0.917	0.938	0.787	0.798	0.927	0.947	0.917	0.947

Table 2. Summary of the final performance of the models on each artifact type defined by the AUROC score.

statistical significance is confirmed by the statistical tests. Evaluation on the Radboud University dataset (Fig. 7) demonstrates an overall improvement, notably for the initially weakest artifact—Air bubbles. Better results are also observed for tissue and focus, with marginal gains for dust and a slight regression for the ink class.

Despite freezing all layers, except the last fully connected layer with 2048 input features and 6 output classes, some overfitting was observed, indicated by an initial increase in loss. However, the loss stabilized over time, and the final performance exhibited less degradation for previously challenging artifacts. The concluding experiment involving a mostly frozen model (Fig. 8) highlights the overfitting issue, with a noticeable performance drop for the dataset with only annotations as a training data. Notably, this regression is absent for the augmented dataset.

Artifact type	$ACR' - ACR$	$ACR'_{anh} - ACR_{anh}$	$RB' - RB$	$RB'_s - RB_s$
Air	- 0.001	0.014	0.107	0.141
Dust	0.069	0.016	0.011	0.009
Tissue	0.114	0.055	0.004	0.005
Ink	- 0.049	- 0.017	- 0.011	0.003
Marker	0.003	- 0.029	- 0.006	- 0.001
Focus	- 0.013	0.024	0.018	0.021
All classes	0.021	0.011	0.020	0.030
p value	1E-03	3E-01	9E-36	4E-06

Table 3. Summary of the improvements in AUROC made by our method in each dataset and for each artifact type. All differences are presented with an additional Wilcoxon signed-rank test performed on an accumulated list of patch predictions.

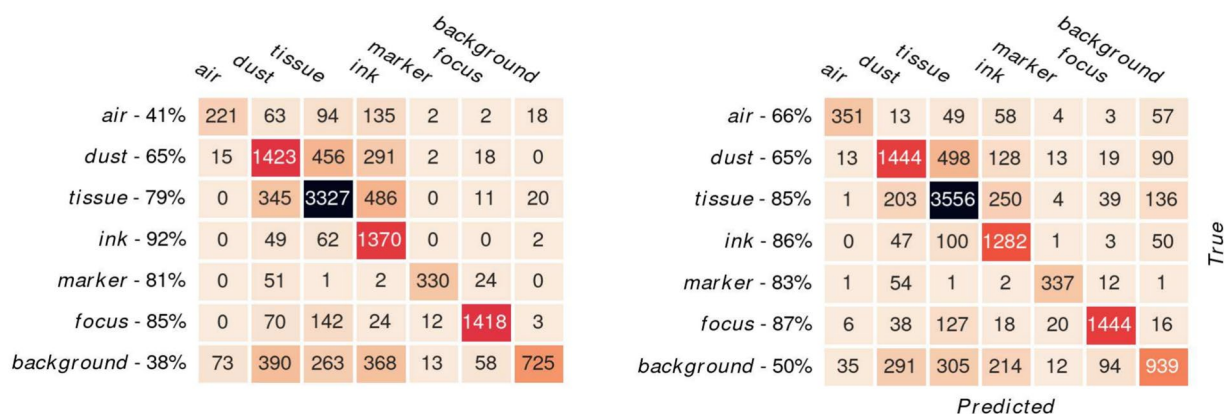


Figure 9. Confusion matrix of predictions after thresholding. (left) Model trained on **RB**. (right) Model trained on augmented **RB'**. Each class displays the percentage of accurate predictions for that specific artifact type.

Comparison between the two sets reveals that the model trained on the augmented dataset outperforms the model trained solely on annotations for all artifact types, indicating the effectiveness of the proposed augmentation approach in mitigating overfitting.

In Fig. 9 we see the confusion matrix after thresholding. Background class was raised when no other class met the required threshold. The high values along the diagonal elements of the matrix indicate that our model was successful in correctly classifying instances across multiple classes. Nonetheless, we have observed patterns in misclassifications, with specific classes exhibiting higher rates of misclassifications, e.g., air, dust, tissue, and background. For each of these classes, our augmentations led to a performance improvement.

Discussion

The presented data augmentation pipeline, blending annotated artifacts from a donor dataset to the destination, demonstrates its utility and positive impact on artifact detection. This method proves effective in generalizing a QC system to real-world histology data, even with limited training annotations, substantially increasing their quantity. Notably, the method excels in handling underrepresented or highly heterogeneous artifacts, evident in significant improvements for artifacts like air and dust. Improvements in handling focus distortions, which are relatively easier to generate correctly, were also observed. The method effectively addresses tissue artifacts, particularly in datasets with similar staining characteristics, though a slight degradation is noted for a complete mix of datasets. This comes from an improved heterogeneity of the training dataset, and could be further improved by increasing the initial sample of the annotated artifacts.

The method's handling of ink artifacts presents a challenge, with potential future improvements in stain transfer methods and the incorporation of advancements from generative deep learning methods. Additionally, the focus artifact type is often mixed with tissue folding. This may be attributed to the frequent presence of both types of artifacts together, as the folds have the potential to cause difficulties with focus. The issue with the background classification could be related to the fact that not all the artifacts present were annotated, and there are patches labeled as containing none of the artifacts when in fact there are.

The study also indicates that the proposed method can mitigate overfitting in high-performing models, as evident in the analysis of validation loss graphs. Quality tissue segmentation significantly influences blur detection, with notable increases in accuracy when proper segmentation is achieved.

Future work could focus on the classification network, exploring potential shortcomings in handling small, irregular shapes or closely situated objects. The study suggests improvements in handling ink artifacts and investigating advanced filtering on the edges of inserted marker artifacts. Further research could explore recent

advancements in generative deep learning methods for stain transfer and expand the artifact types, requiring more collaboration with professional pathologists and incorporating additional datasets for a more diverse artifact collection.

Methods

Overview

The proposed framework addresses the challenge of augmenting Whole Slide Images (WSIs) with artifacts through an optimized processing pipeline. To efficiently handle large-scale histopathology datasets, the framework adopts a streaming approach for augmentation, loading patches sequentially to overcome memory constraints. Images are saved in the *.tiff* format, compatible with the standard pyramidal image structure of OpenSlide—an open standard in histopathology, ensuring compatibility with prevalent tools^{22,23}. Annotations from the artifact detection framework are stored in the *.xml* format, chosen for its human-readable nature and compatibility with ASAP (Automated Slide Analysis Platform) software²⁴.

To leverage high-resolution capabilities without downsampling, the framework strategically reads only the Region of Interest (ROI) instead of the entire image. Focusing solely on areas with artifacts maximizes resolution utilization, ensuring no loss of critical details during the detection process. The pipeline begins by referencing the image and corresponding annotations, with the image serving as the source for artifact extraction and annotations providing vital information. This iterative extraction process establishes the groundwork for subsequent processing steps, ensuring accuracy in artifact identification.

Following the extraction of artifacts from the Region of Interest (ROI), the pipeline stores each artifact in a dedicated collection, referred to as the Artifact Collection. This repository retains all detected artifacts, serving as a reference for subsequent stages. The augmentation process, illustrated in Fig. 10, iteratively blends these artifacts onto the destination WSI. Throughout this augmentation process, the following sequential steps are executed:

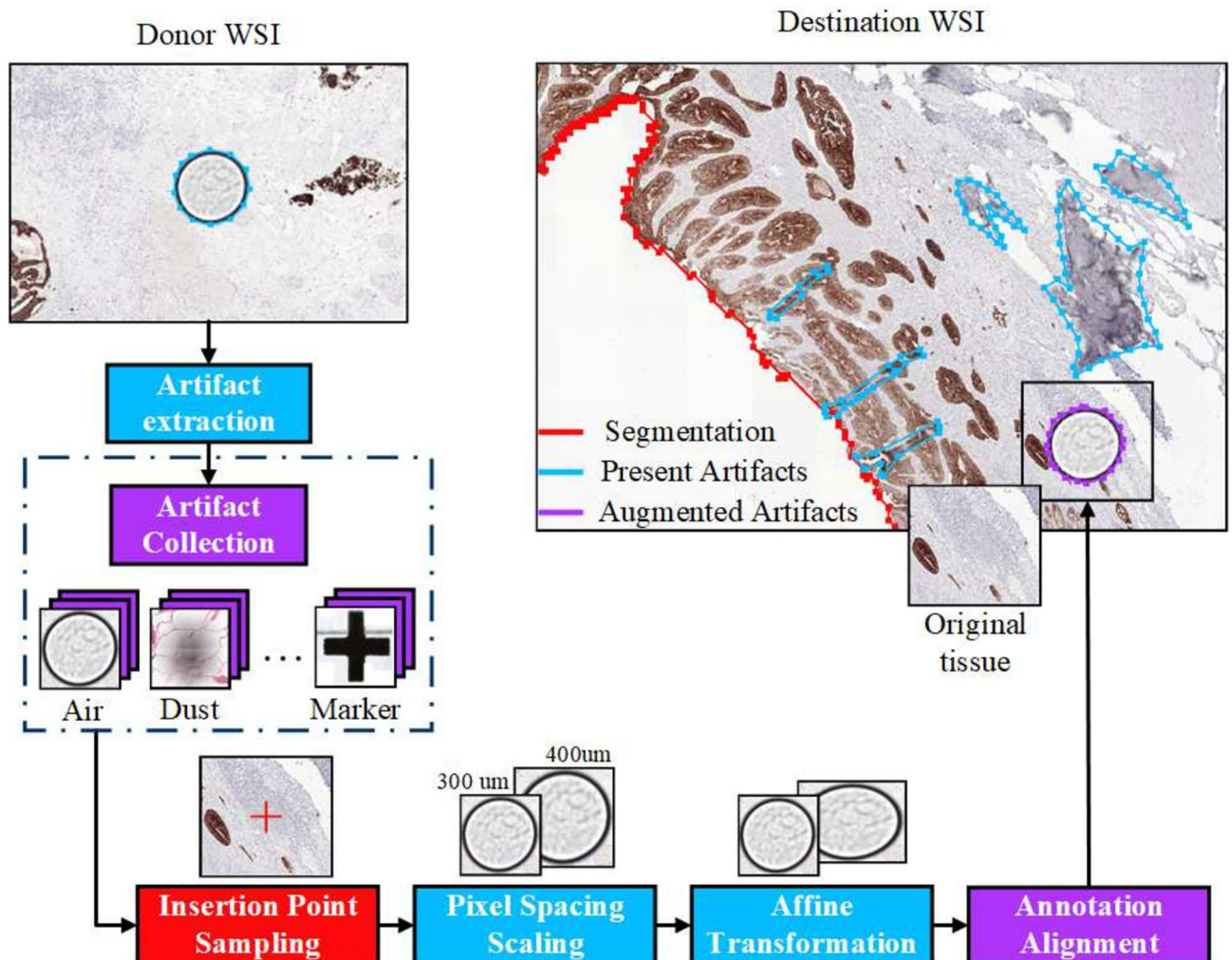


Figure 10. The process of augmenting a WSI with an artifact coming from previously extracted artifact collection.

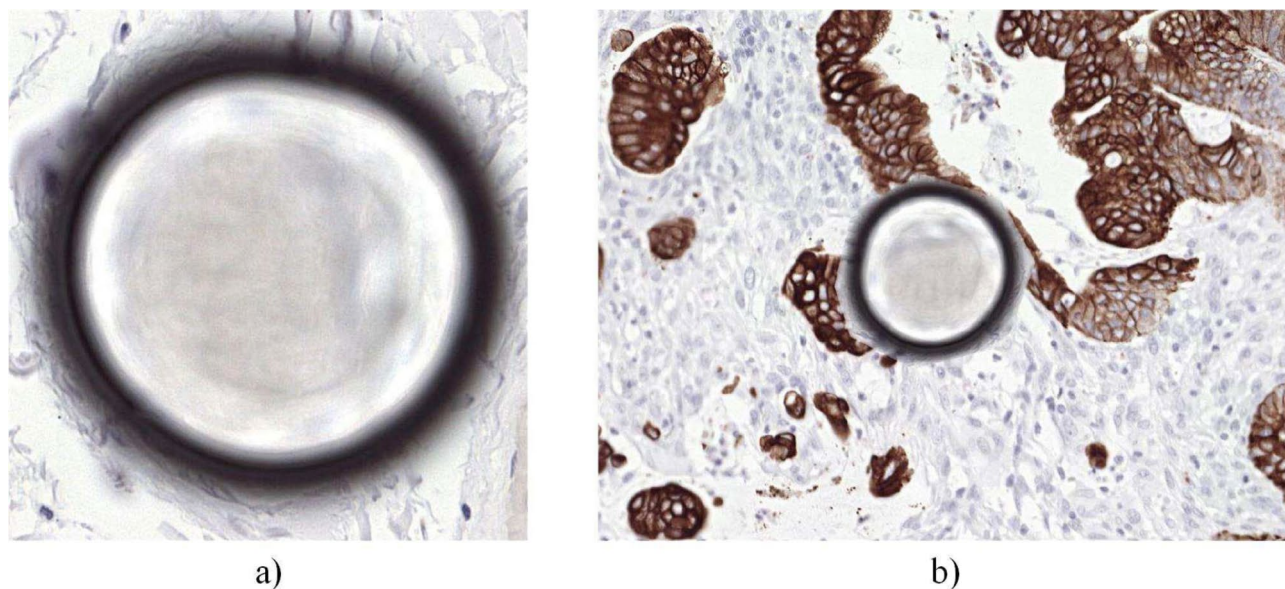


Figure 11. Example of an air bubble being blended onto the destination WSI: (a) original image, (b) artifact blended with the destination image.

Artifact type	Max no. of inserted artifacts	Location of inserted artifacts
Air	4	Whole WSI
Dust	7	Whole WSI
Tissue	4	On top of the tissue (foreground)
Ink	4	Edge of the tissue
Marker	4	Outside the tissue (background)
Focus	2	On top of the tissue (foreground)

Table 4. Summary of the used configuration for each artifact type.

- Sampling of the artifact insertion point: The central point for artifact insertion is determined based on the specific artifact type, with detailed configurations provided in Table 4.
- Scaling the artifact to correct pixel spacing: This critical step adjusts artifacts originating from different scanners or tissues to maintain real-life measurements. File metadata is utilized to preserve the physical measurements of each WSI's pixels.
- Affine transformation: To diversify the resulting data, artifacts undergo augmentation, incorporating randomly applied rotation and scaling. Affine transformation is selected for its avoidance of tearing or folding artifacts.
- Aligning the annotation: Following all applied operations on the artifact image, its annotation undergoes an identical transformation to ensure alignment with the augmented artifact.

To address the necessity for stain-invariant segmentation for insertion point sampling, we used a custom deep learning model. For detailed architecture and training pipeline specifics, readers are directed to our previous paper²⁵. The method was selected after a thorough consideration of available options, acknowledging the critical need for a stain-invariant model capable of handling images coming from different datasets with diverse magnifications. This stain-invariant and multiresolution segmentation increases the overall robustness of the framework across different datasets and staining methodologies.

$$I_{out} = Gauss(M_{artifact}) \times I_{artifact} + (1 - Gauss(M_{artifact})) \times I_{WSI} \quad (1)$$

Moreover, the pipeline can cut the augmented WSI into patches. Leveraging the segmentation module, this allows for even sampling of each artifact type, while simultaneously preserving a specified number of empty, background patches. This balanced training approach contributes to fewer false positive predictions. To execute the blending process, distinct strategies are employed for each artifact group:

- Focus distortions (Gaussian blurring): Gaussian blurring simulates focus distortions in the blending process.

- Markers and air bubbles: These artifacts are blended by insertion onto the destination image, followed by bilateral filtering.
- Dust artifacts: Investigated with seamless cloning (gradient editing) and the previously described method for markers.
- Ink transfer: Reinhard Color Normalization²⁶ is employed to maintain the original tissue structure during ink transfer. Alternative methods such as²⁷ were explored, but did not yield superior results and increased computational costs.

After each blending step, further processing is applied to the edges of the inserted artifact. A smooth transition, described by Eq. (1), is implemented. This equation outlines a gradual, linear transition between pixel values of the artifact and the destination image based on a Gaussian-smoothed artifact's annotation mask. Figure 11 visually depicts artifacts both before and after the blending process, showcasing the efficacy of the proposed methodology.

Classification

Image patches were classified to evaluate potential challenges, especially for small, irregular shapes or closely situated objects²⁸. Patches with artifacts and backgrounds of size 224 × 224 px were cut from the images.

A pretrained ResNet50 model was selected for this classification task. In most experiments, the whole model was trained. In the case of **RB_s** and **RB_s'** datasets, the last convolutional layer and one fully connected layer at the end of the network were unfrozen to mitigate overfitting. The model's feasibility was assessed by training it on dataset annotations and then on an augmented dataset. Two evaluations were conducted: one with additional unseen annotation patches and another with annotation patches from an entirely new dataset to assess generalizability. Two additional experiments utilized professionally annotated data, training the model on annotations only and on an augmented dataset that included data from all three datasets. The experiment was evaluated on an unseen subsample of the professional annotations. The last experiment was repeated with a reduction of trainable model layers to only the last fully connected layer to assess potential overfitting.

Ethical approval

For this study we are using existing data coming from two open datasets (ACROBAT and ANHIR) which received ethical approval from their respective institutional review boards. For data coming from the Radboud University, all methods were carried out in accordance with relevant guidelines and regulations. Experimental protocols and patient inclusion were approved by the ethical review board (METC) of Radboudumc, Nijmegen, Netherlands. For all included patients, informed consent was obtained.

Conclusions

In conclusion, our QC system demonstrates efficacy in accurate artifact detection and classification in histopathology images. Utilizing annotated datasets with advanced data augmentation techniques promises improved performance and reduced overfitting. While challenges in handling specific artifacts and adapting to various magnification levels and datasets exist, this study contributes to developing a reliable histopathology QC system for enhanced image analysis and accurate clinical diagnosis. We release the code freely at²⁹.

Data availability

The repository²⁹ contains the code, model weights, and manual data annotations. Training data is sourced from^{19,20}, both of which offer open-source datasets. Additional data is provided by Radboud University as part of the BigPicture project, and the Radboud dataset will be accessible on the BigPicture platform.

Received: 23 February 2024; Accepted: 26 July 2024

Published online: 01 August 2024

References

1. Brixel, R. *et al.* Whole slide image quality in digital pathology: Review and perspectives. *IEEE Access Pract. Innov. Open Solut.* **10**, 131005–131035 (2022).
2. Khan, S., Tijare, M. S., Jain, M. & Desai, A. Artifacts in histopathology: A potential cause of misinterpretation. *Res. Rev. J. Dent. Sci.* (2014).
3. Elias, J. M. *et al.* Special report: Quality control in immunohistochemistry: Report of a workshop sponsored by the biological stain commission. *Am. J. Clin. Pathol.* **92**, 836–843. <https://doi.org/10.1093/ajcp/92.6.836> (1989).
4. Tsutsumi, Y. Pitfalls and caveats in applying chromogenic immunostaining to histopathological diagnosis. *Cells* **10**, 1501. <https://doi.org/10.3390/cells10061501> (2021).
5. Taqi, S. A., Sami, S. A., Sami, L. B. & Zaki, S. A. A review of artifacts in histopathology. *J. Oral Maxillofac. Pathol. JOMFP* **22**, 279. https://doi.org/10.4103/jomfp.JOMFP_125_15 (2018).
6. Ekundina, V. & Eze, G. Common artifacts and remedies in histopathology (a review). *Afr. J. Cell. Pathol.* **4**, 6–12. <https://doi.org/10.5897/AJCPATH15.002> (2015).
7. Kanwal, N., Perez-Bueno, F., Schmidt, A., Engan, K. & Molina, R. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access* **10**, 58821–58844. <https://doi.org/10.1109/ACCESS.2022.3176091> (2022).
8. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inform.* **3**, 1–7. <https://doi.org/10.1200/CCI.18.00157> (2019).
9. Chen, Y. *et al.* Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. *J. Pathol.* **253**, 268–278. <https://doi.org/10.1002/path.5590> (2021).
10. choosehappy. Choosehappy/HistoQC. <https://github.com/choosehappy/HistoQC/wiki> (2024).

11. Campanella, G. *et al.* Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Comput. Med. Imaging Graph.* **65**, 142–151 (2017).
12. Senaras, C., Niazi, M. K. K., Lozanski, G. & Gurcan, M. N. DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. *PLoS One.* **13**.
13. Babaie, M. & Tizhoosh, H. R. Deep features for tissue-fold detection in histopathology images. In *Digital Pathology* (eds. Reyes-Aldasoro, C. C., *et al.*), 125–132. https://doi.org/10.1007/978-3-030-23937-4_15 (Springer International Publishing, 2019).
14. Maleki, D., Afshari, M., Babaie, M. & Tizhoosh, H. R. Ink marker segmentation in histopathology images using deep learning. In *Advances in Visual Computing* (eds. Bebis, G. *et al.*) 359–368. https://doi.org/10.1007/978-3-030-64556-4_28 (Springer International Publishing, 2020).
15. Smit, G. & Cigéhn, M. Quality control of whole-slide images through multi-class semantic segmentation of artifacts.
16. Foucart, A., Debeir, O. & Decaestecker, C. Artifact identification in digital pathology from weak and noisy supervision with deep residual networks. In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)* 1–6. <https://doi.org/10.1109/CloudTech.2018.8713350> (IEEE, 2018).
17. Schömig-Markiefka, B. *et al.* Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod. Pathol.* **34**, 2098–2108.
18. Wang, N. C. *et al.* Stress testing pathology models with generated artifacts. *J. Pathol. Inform.* **12**.
19. Weitz, P. *et al.* The challenge: Automatic registration of breast cancer tissue. In *2022 Medical Image Analysis*, vol. 103257. <https://doi.org/10.1016/j.media.2024.103257> (2024) <https://www.sciencedirect.com/science/article/pii/S1361841524001828>.
20. Borovec, J. *et al.* ANHIR: Automatic non-rigid histological image registration challenge. *IEEE Trans. Med. Imaging* **39**, 3042–3052. <https://doi.org/10.1109/TMI.2020.2986331> (2020).
21. Litjens, G. *et al.* 1399 h & e-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience.* **7**, giy065. <https://doi.org/10.1093/gigascience/gy065>.
22. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27. <https://doi.org/10.4103/2153-3539.119005>.
23. Goode, A. & Satyanarayanan, M. A vendor-neutral library and viewer for whole-slide images. In *Computer Science Department, Carnegie Mellon University, Technical Report CMU-CS-08-136*.
24. Computationalpathologygroup/ASAP. <https://github.com/computationalpathologygroup/ASAP> (Computational Pathology Group, 2024).
25. Jurgas, A., Wodzinski, M., Atzori, M. & Müller, H. Robust multiresolution and multistain background segmentation in whole slide images. In *The Latest Developments and Challenges in Biomedical Engineering, Lecture Notes in Networks and Systems* (eds. Strumillo, P., Klepaczko, A., Strzelecki, M. & Bociąga, D.) 29–40. https://doi.org/10.1007/978-3-031-38430-1_3 (Springer Nature Switzerland).
26. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41. <https://doi.org/10.1109/38.946629> (2001).
27. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. vol. 9, 1107–1110. <https://doi.org/10.1109/ISBI.2009.5193250>.
28. Guo, Z., Wang, C., Yang, G., Huang, Z. & Li, G. MSFT-YOLO: Improved YOLOv5 based on transformer for detecting defects of steel surface. *Sensors* **22**, 3467. <https://doi.org/10.3390/s22093467> (2022).
29. Jurgas, A. Jarartur/HistopathologyAugmentationResearch. <https://github.com/Jarartur/HistopathologyAugmentationResearch> (2024).

Acknowledgements

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945358. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA, Belgium (www.imi.europa.eu). The research reflects only the author's view and the Joint Undertaking is not responsible for any use that may be made of the information it contains. Additionally, the research was supported in part by PLGrid Infrastructure. We gratefully acknowledge Poland's high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016239.

Author contributions

A.J. and M.W. conceptualized the method, A.J. carried out the implementation, conducted the experiments and evaluation. Data acquisition was done by M.A. (Marina D'Amato) and J.L. Results analysis was done by all authors. The article was prepared by A.J. and M.W. and reviewed by all the authors. J.L., M.A. (Manfredo Atzori), and H.M. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.J.

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024