TROPIC: a Training-Free Framework for Patch-Level Pathology Quality Control

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. The digitization of pathological slides involves both manual and machine processes. These processes often introduce artifacts that degrade image quality, posing challenges for clinical diagnosis and AI model training. Existing quality control methods are either too slow to process large datasets or inaccurate for downstream model training. To achieve efficient quality control without sacrificing model performance, we propose TROPIC: a Training fRee framewOrk for Pathology qualIty Control. Briefly, our framework constructs a reference set and employs a accelerated post-processing voting mechanism to predict artifact labels. It effectively identifies common artifacts, such as tissue folds, out-of-focus areas, and air bubbles in whole slide images (WSIs). Compared to existing classification based methods, our approach improves F1-score by over 10% with 1.14× speed-up. Furthermore, compared with the existing SOTA segmentation-based method, it achieves a 14× speed-up while maintaining downstream model performances. Additionally, it achieves 100% accuracy on filtering out non-pathological images in public pathology datasets. Our framework are opensourced at https://github.com/user98377861/TrainFree .

Keywords: Quality Control · Data Pruning · Dataset Cleaning · Training-Free.

1 Introduction

Manual procedures (such as tissue processing, sectioning, and staining) and scanning of pathology slides introduce artifacts that lead to image degradation. Thus, slide quality control (QC) has been a long-term problem in clinical diagnosis and computational pathology. Recently, pathology foundation models such as [5,13] have shown superior performance and great potential for most computational pathology tasks. These foundation models must be trained on unprecedented large-volume datasets, making a fast and effective QC method an essential component of modern pathology data processing.

Several pathological QC methods have been proposed recently. HistoQC [12] employs non-deep learning techniques for WSI-level quality control, eliminating the need for manual labeling but being outperformed by later deep learning methods. HistoROI [14] uses active learning to train a classification model for

one artifact type but its performance is limited. PathProfiler [8] trains a multi-task deep neural network for comprehensive artifact detection but is limited to a single diagnostic domain. Other works ([17], [10], [3], and [4]) also focus on single artifact types, restricting their practical utility. The most recent GrandQC [21] trains pixel-level semantic segmentation models, identifying 7 kinds of WSI artifacts with superior performance. However, it is almost not applicable for large-scale patch-level processing due to the high computational cost of segmentation models.

Noise removal methods like [2] and [19] developed classifiers for specific noise in datasets crawled from social media (such as Quilt-1M [11]). They are either resource-intensive because of manual labeling and extensive training or not generally adequate due to the limited data distribution.

To summarize, these existing QC methods face two key issues: (1) segmentation-based approaches are accurate but slow, whereas classification-based methods are fast but less precise; and (2) trivial classification models can hardly handle non-pathological images from the public domain with tremendous classes. Aiming to solve these issues, we propose TROPIC: a Training fRee framewOrk for Pathology qualIty Control, which provides general high-efficiency patch-level QC without sacrificing accuracy.

To achieve high-efficiency, we design a training-free framework to classify given patches. To be specific, we first construct a reference set using existing datasets or segmentation models, and then identify each test sample's top K nearest neighbors from the reference set. The class with the most votes among these K samples is selected as the predicted result. This gets rid of the slow segmentation models but brings a following challenge: how to accelerate the top-K searching when reference sets are large. Therefore, we also build a retrieval approach that utilizes GPU-accelerated approximate nearest neighbor (ANN) for real-time similarity top-K voting.

The top-K voter naturally serves as a general classifier when reference sets are extended. Therefore, to achieve the generalized filtering for tremendous public domain classes, we utilize public datasets from the computer vision community (such as [6]) to build a large and general reference set for non-pathology images. We apply it to filter out non-pathological images in a mixed dataset composed of public pathology datasets [19] and general datasets [18], achieving 100% accuracy.

The main contributions of our work are as follows:

- We introduce a training-free framework for patch-level quality control of pathological images.
- We enhance the efficiency of our framework by integrating a retrieval-based method.
- We apply our framework to two practical scenarios, showcasing its competitive performance: (1) cleaning large-scale pathological datasets by filtering out non-pathological images, and (2) enabling patch-level quality control in the foundation model training process.

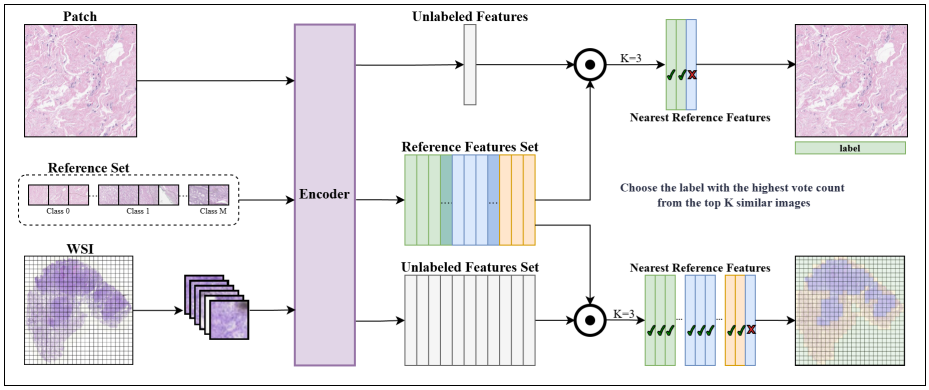


Fig. 1: Framework of our method. We have two kinds of inputs: WSI and patch. There are three stages: i) reference set and test set are encoded by a same pretrained encoder; ii) test set is compared with the whole reference set and top K neareset neighbors are chosen; iii) these neighbors vote for the final prediction.

2 Method

Our framework (Fig.1) for the quality control task accepts two kinds of input: patches and WSIs. For the first type of input, the corresponding label can be directly returned by the framework. For the second type of input, the method first divides the WSI into 512×512 patches, then compares them with the reference set, and outputs the corresponding labels for each patch of the entire WSI. As for the data pruning task, it only accepts patch-level pathological images as input and outputs a Boolean label representing the class.

For the two distinct tasks, we employ different methodologies to construct the reference sets. Regarding the patch-level quality control task, we construct the reference set using GrandQC's test dataset. It is a manually labeled segmentation dataset, but we can convert this dataset into a classification dataset based on pixels' label ratios. Similarly, after encoding, we obtain the desired feature reference set necessary for our analysis. In the context of the data pruning task, we utilize the CC3M [18] (a general-domain multimodal dataset) and PathCap [19] (a pathology multimodal dataset) datasets to establish our reference sets. We select one hundred thousand images from each dataset to form the reference sets, which are subsequently encoded using an encoder to obtain the feature reference sets for comparison.

The framework operates by identifying the top K nearest neighbors within the reference set and determining the final result based on the most frequent label among these neighbors. Formally, let the reference set be denoted as $D_{ref} = \{(x_i, y_i)\}_{i=1}^{M}$, where $x_i \in R^{H \times W \times C}$ represents the input patches and $y_i \in R$ denotes their corresponding labels. The number of patches of the reference set is M. A pretrained image encoder G is employed to generate embeddings $\{q_i\}_{i=1}^{N}$ for the patches, where $q_i \in R^{d_G}$ and d_G is the embedding dimension determined

by G. For a test patch x_{test} from the test set D_{test}, its embedding q_{test} is similarly derived.

The framework proceeds by computing the similarity between q_{test} and all embeddings in the reference set. The top K samples with the highest similarity scores are selected, as formalized in Eq. 1. Finally, the labels $\{y_i\}_{i=1}^{K}$ of these K nearest neighbors are aggregated, and the label with the highest frequency is assigned as the final prediction, as formalized in Eq. 2. This approach ensures robust and interpretable results by leveraging top K nearest neighbors.

$$I_{top_K} = \arg \max_{S \subseteq \{q_1, q_2, \ldots, q_M\}, |S|=K} q_{test} q_i \tag{1}$$

$$y_{test} = \arg \max_{y_i} (Counter(y_i)), i \in I_{top_K} \tag{2}$$

The top K approach assumes that a correctly labeled sample is more similar to multiple samples in its true category than to those in other categories. This strategy enhances noise robustness: relying on a single highest-similarity label is error-prone under label noise, as outliers can influence results. Fig. 2 empirically validates this, showing that increasing K improves classification accuracy by aggregating information from multiple neighbors, mitigating label noise. However, performance does not improve indefinitely with K. For example, when $K = M$, the entire reference set determines the result through voting, causing the most frequent class to dominate predictions regardless of the test sample's true characteristics. Thus, performance initially improves with K but declines after exceeding an optimal threshold. A small K may miss sufficient accurate samples, while an excessively large K over-relies on global class proportions, diluting accurate samples' influence. This underscores the importance of selecting an appropriate K. Empirically, we set $K = 3$ as a conservative yet effective choice.

The reference set size significantly influences predictions: a larger, accurately labeled set enhances classification accuracy by providing richer information. This is empirically supported in Fig. 3, showing a positive correlation between reference set size and performance metrics. While a small reference set sufficed for our relatively simple task, more complex scenarios with additional categories would require proportionally larger reference sets.

3 Experiments

3.1 Datasets

For the data pruning task, we employ two domain-specific datasets: a general-domain image dataset and a pathology-specific image dataset. The general-domain dataset is constructed using CC3M and the pathology image dataset is constructed using PathCap. The PathCap contains 207,000 high-quality pathology image-caption pairs collected from PubMed and internal pathology guidelines books. In the experiment, we benchmarked our method against alternative

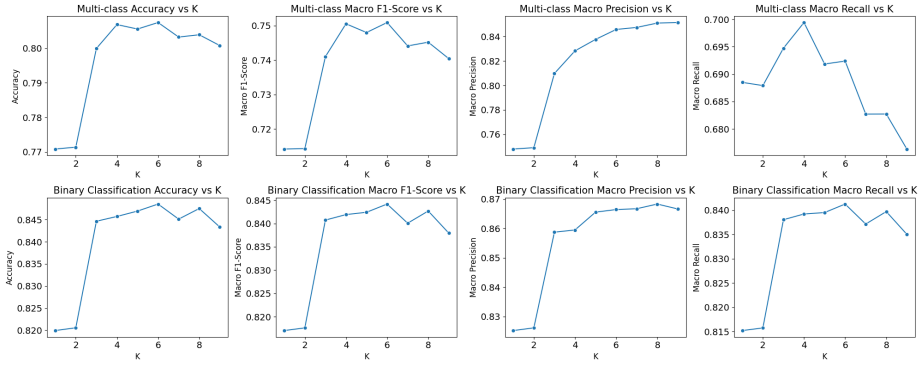


Fig. 2: The influence of different K values on binary classification and multi-label classification in the quality control task.

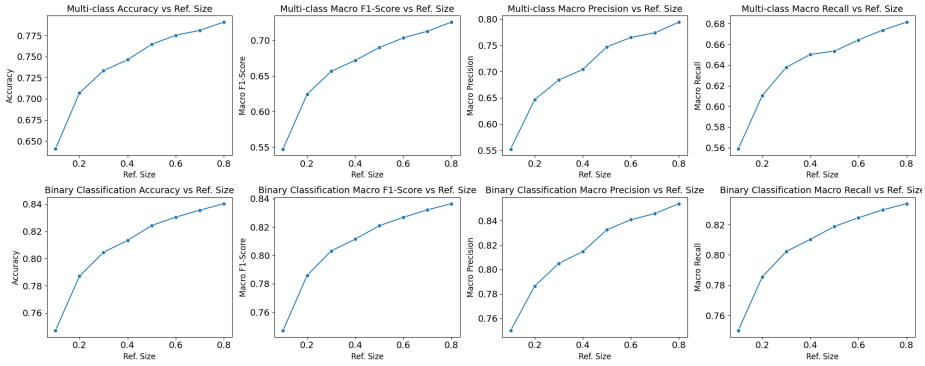


Fig. 3: The influence of different reference sizes on binary classification and multi-label classification in the quality control task. Ref. Size on the x-axis represents the ratio of the experimental set size to the original set size.

binary classifiers in the data pruning task through binary classifiers trained on 200,000 CC3M samples and 200,000 TCGA-derived patches. To account for potential variations in accuracy across different architectures and parameter scales, we trained four distinct types of classifiers: MobileNetV2 [16], EfficientNet-B3 [20], ResNet-50 [9], and ViT-B/32 [7]. This selection encompasses a diverse range of model complexities, from lightweight convolutional networks to transformer-based architectures, ensuring a comprehensive evaluation of performance. Evaluation was performed on ten 100,000-sample datasets randomly drawn from CC3M's unused portion and ten 100,000-sample datasets from PathCap.

For the quality control task, we utilize GrandQC's open-source, manually labeled segmentation dataset. Patch-level classification is derived from pixel label distribution ratios. To construct the reference set, we randomly sample 20% of the data for testing and use the remainder for reference. This split is repeated

Table 1: Quality control task's performance comparisons with HistoROI. Classes are simplified to binary: normal patch and patch with artifacts.

Model	Class w/o artifacts		Class artifacts		Macro Avg.		Time Usage(s)	
	HistoROI	Ours	HistoROI	Ours	HistoROI	Ours	HistoROI	Ours
Breast	58.91%	77.81%	58.89%	89.15%	58.90%	83.48%	3.27	1.62
Colon	69.22%	82.94%	57.76%	84.94%	63.49%	83.94%	3.63	2.35
Kidney	81.92%	89.65%	39.23%	73.38%	60.58%	81.51%	3.50	3.35
Prostate	73.43%	86.74%	38.77%	74.34%	56.10%	80.54%	3.68	2.22
total	74.13%	86.57%	49.73%	81.58%	61.93%	84.07%	10.88	9.52

Table 2: Quality control task's performance comparisons with PathProfiler. Original classes are adjusted to four classes.

Model	Precision		Recall		F1	
	PathProfiler	Ours	PathProfiler	Ours	PathProfiler	Ours
Class w/o Arti.	69.93%	79.11%	2.04%	94.86%	3.98%	86.27%
Class FOLD	97.96%	74.98%	4.76%	57.38%	9.07%	65.00%
Class OOF	15.75%	86.39%	66.78%	73.61%	25.45%	79.49%
Class Other Arti.	32.75%	85.65%	62.40%	63.17%	42.96%	72.71%
Macro Avg.	54.10%	81.53%	33.99%	72.26%	20.37%	75.87%

ten times to reduce performance uncertainty. For experiments involving our local private dataset to train a UNI-style model, we employ a reference set of 100,000 patches generated by GrandQC's segmentation model (mpp=1) and evaluate on BACH [1] dataset .

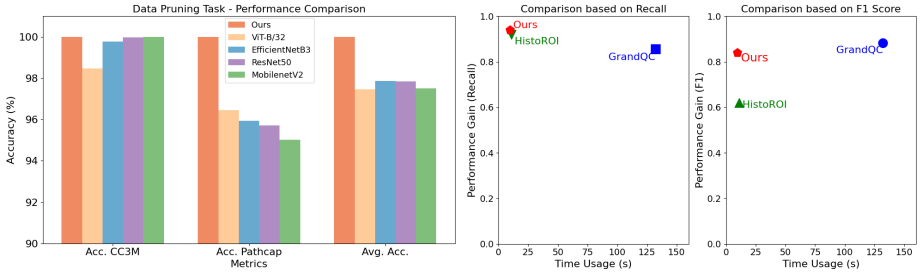
3.2 Implementation details

For the data pruning task, all models were trained with a batch size of 256 and a learning rate of 0.0001, utilizing 8 A800 GPUs for efficient training. In the quality control task, we employed the image encoder from CLIP (ViT-B/16) [15] to generate patch embeddings. For time efficiency comparisons, all baseline models were evaluated with a consistent batch size of 400. Notably, all reported time measurements were exclusively calculated during the inference phase to ensure a fair and accurate assessment of computational performance.

3.3 Results

We use accuracy to evaluate the data pruning task, with results shown in Fig. 4a. While other models achieve high accuracy (>99%) on natural images, their performance declines on pathological images. In contrast, our method achieves 100% accuracy on both domains, demonstrating superior and consistent effectiveness.

We first compare our method with GrandQC. As shown in Fig. 4b, our approach achieves significantly faster processing while maintaining high macro F1



(a) Performance comparison of different models for data pruning task.

(b) Performance comparison of different methods for quality control task.

Fig. 4: (a) Non-pathological filtering task's performance across classes; (b) Performance gain over time with efficiency comparison.

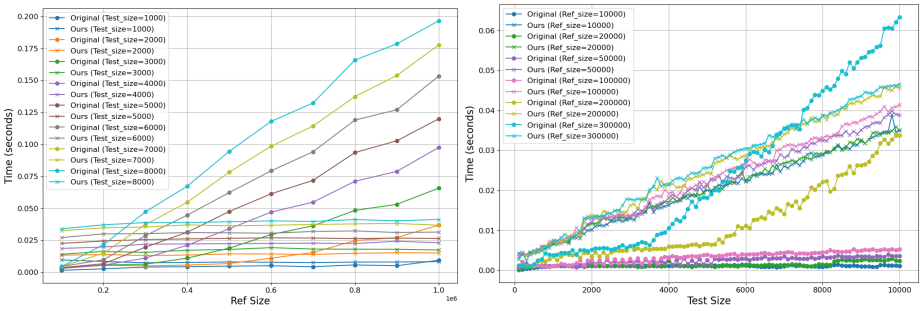
Table 3: Performance of UNI-style model trained on local dataset : linear probing (LP), KNN probing (KNN), and nearest class prototype (NCP). The numbers in brackets represent the number of patches in the training set, where 0.79M is the unfiltered dataset and 0.76M is the filtered dataset.

Model	LP		KNN		NCP	
	Accuracy	Weighted F1	Accuracy	Weighted F1	Accuracy	Weighted F1
UNI	86.7%	86.6%	80.0%	79.7%	75.0%	74.6%
Ours(0.79M)	88.3%	87.7%	79.2%	76.9%	85.0%	84.2%
Ours(0.76M)	90.8%	90.4%	80.8%	78.1%	85.8%	84.4%

score and recall. Next, we compare our method with two baseline approaches: HistoROI and PathProfiler. Unlike segmentation-based methods like HistoQC and GrandQC, HistoROI and PathProfiler employ classification approaches, making them more suitable for direct speed and accuracy comparisons. To align with their artifact categorization, we adjust the reference dataset labels for separate evaluations. For HistoROI, we report the macro F1 score to assess performance across organs, as shown in Table 1. For PathProfiler, due to its precision-recall imbalance, we present both metrics individually along with the macro F1 score for a comprehensive evaluation, as detailed in Table 2.

We also evaluated the impact of removing artifact-containing patches on visual model training. Using our local dataset of 3484 breast cancer WSIs, we trained a UNI-style [5] model. As shown in Table 3, removing artifact-containing patches improves accuracy on the BACH dataset, evaluated using multiple methods.

As previously mentioned, matrix multiplication significantly slows down nearest neighbor retrieval. To address this, we simulated a scenario involving larger datasets, including expanded reference sets and test sets. Our experiments confirm that incorporating CARGA accelerates retrieval, enabling the method to maintain high efficiency even as the dataset size grows.



(a) Different Reference Sizes (b) Different Test Sizes

Fig. 5: Time cost statistics. (a) simulates performance under different sizes of reference set. (b) simulates performance under different sizes of test set.

Fig. 5a shows that, for fixed test set sizes, our method maintains nearly constant computation time as the reference set grows, unlike the original method, which suffers from significantly increased computation time due to matrix multiplication. Fig. 5b further demonstrates that our method scales nearly linearly with test set size, while the original method experiences a rapid complexity increase when total data volume exceeds a threshold. These results underscore our method's superior scalability and efficiency for large-scale data retrieval, particularly in real-time pathological image quality assessment.

4 Conclusions

In this work, we introduce a novel training-free method for patch-level pathological image quality control and data pruning. Through extensive experimentation, we systematically analyze the impact of key hyperparameters, including the selection of K and the size of the reference set, on model performance. Our results demonstrate the effectiveness of the proposed method across both quality control and data pruning tasks, as validated by comprehensive empirical evaluations. We also propose retrieval solutions for possible complex tasks. In the future, we plan to investigate its application in multiple instance learning.

References

1. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical image analysis **56**, 122–139 (2019)
2. Aubreville, M., Ganz, J., Ammeling, J., Kaltenecker, C., Bertram, C.: Model-based cleaning of the quilt-1m pathology dataset for text-conditional image synthesis. In: Medical Imaging with Deep Learning

3. Bautista, P.A., Yagi, Y.: Improving the visualization and detection of tissue folds in whole slide images through color enhancement. Journal of pathology informatics **1**(1), 25 (2010)
4. Campanella, G., Rajanna, A.R., Corsale, L., Schüffler, P.J., Yagi, Y., Fuchs, T.J.: Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. Computerized medical imaging and graphics **65**, 142–151 (2018)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
8. Haghighat, M., Browning, L., Sirinukunwattana, K., Malacrino, S., Khalid Alham, N., Colling, R., Cui, Y., Rakha, E., Hamdy, F.C., Verrill, C., et al.: Automated quality assessment of large digitised histology cohorts by artificial intelligence. Scientific Reports **12**(1), 5002 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hosseini, M.S., Brawley-Hayes, J.A., Zhang, Y., Chan, L., Plataniotis, K.N., Damaskinos, S.: Focus quality assessment of high-throughput whole slide imaging in digital pathology. IEEE transactions on medical imaging **39**(1), 62–74 (2019)
11. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology (2023)
12. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A.: Histoqc: an open-source quality control tool for digital pathology slides. JCO clinical cancer informatics **3**, 1–7 (2019)
13. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
14. Patil, A., Diwakar, H., Sawant, J., Kurian, N.C., Yadav, S., Rane, S., Bameta, T., Sethi, A.: Efficient quality control of whole slide pathology images with human-in-the-loop training. Journal of Pathology Informatics **14**, 100306 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
17. 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**(10), e0205387 (2018)

18. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
19. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Shui, Z., Yu, X., Zhao, Y., Li, H., Zhang, Y., Zhao, R., et al.: Pathasst: Redefining pathology through generative foundation ai assistant for pathology. arXiv preprint arXiv:2305.15072 **2** (2023)
20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
21. Weng, Z., Seper, A., Pryalukhin, A., Mairinger, F., Wickenhauser, C., Bauer, M., Glamann, L., Bläker, H., Lingscheidt, T., Hulla, W., et al.: Grandqc: A comprehensive solution to quality control problem in digital pathology. Nature Communications **15**(1), 1–12 (2024)