Pathological image binary classifier

WENJIN QI

March 17, 2025

Abstract

Currently, most publicly available large-scale multimodal pretraining datasets in pathology are created from Internet. Despite efforts to curate these datasets, a significant number of non-pathological images remain. These non-pathological images are often inaccurately paired with pathological descriptions, leading to mismatched image-text pairs that adversely affect the training of multimodal models and degrade their performance. To address this issue, we propose a filtering approach using a binary classifier trained on patches extracted from Whole Slide Images (WSIs) spanning 15 classes from The Cancer Genome Atlas (TCGA) and general images from Conceptual captions 3M (CC3M) [23]. This classifier effectively identifies and removes non-pathological images. Our results demonstrate that datasets filtered using this model achieve better finetuning performance compared to finetuning on unfiltered datasets. Furthermore, we evaluate the performance of various models as filtering classifiers and find that the best-performing model, UNI, filters out more than **60%** of the dataset while still yielding a **3%** improvement in downstream tasks.

1 Introduction

The success of current multimodal large models, such as CLIP [21], can be largely attributed to the availability of large-scale multimodal datasets like LAION-400M [22], which are primarily obtained through web scraping. Inspired by these advances, researchers have started curating pathological multimodal datasets (e.g., Quilt-1M [14]) to accelerate the development of multimodal models tailored for pathology. These datasets are often sourced from diverse platforms, including YouTube, Twitter, research articles, and general internet searches. However, datasets derived from web scraping present substantial challenges that must be addressed. One of the most critical issues is the inclusion of a large number of non-pathological images inaccurately paired with pathology-related text. This not only results in the contamination of datasets with irrelevant images but also introduces significant errors in image-text correspondence. Such noise undermines the quality of the data and often leads to degraded model performance, as demonstrated in prior work on the impact of noisy correspondences [13]. Given the growing reliance on web-scraped datasets for pathological multimodal research, it is imperative to develop effective strategies to mitigate these issues. Addressing this challenge is crucial for ensuring that the resulting models are robust, accurate, and capable of driving advancements in pathology-focused applications.

2 Related Work

Multimodal Data Pruning The pruning of image-text datasets has been extensively studied in general domains, with existing methods broadly categorized into two types: those that leverage external knowledge and those that operate solely within the given dataset and model framework. As Datacomp [10] has already became a standard benchmark to compare different data pruning methods, we refer datasets and model specified in Datacomp as given knowledge. Any additional data or models not included in Datacomp are regarded as external knowledge.

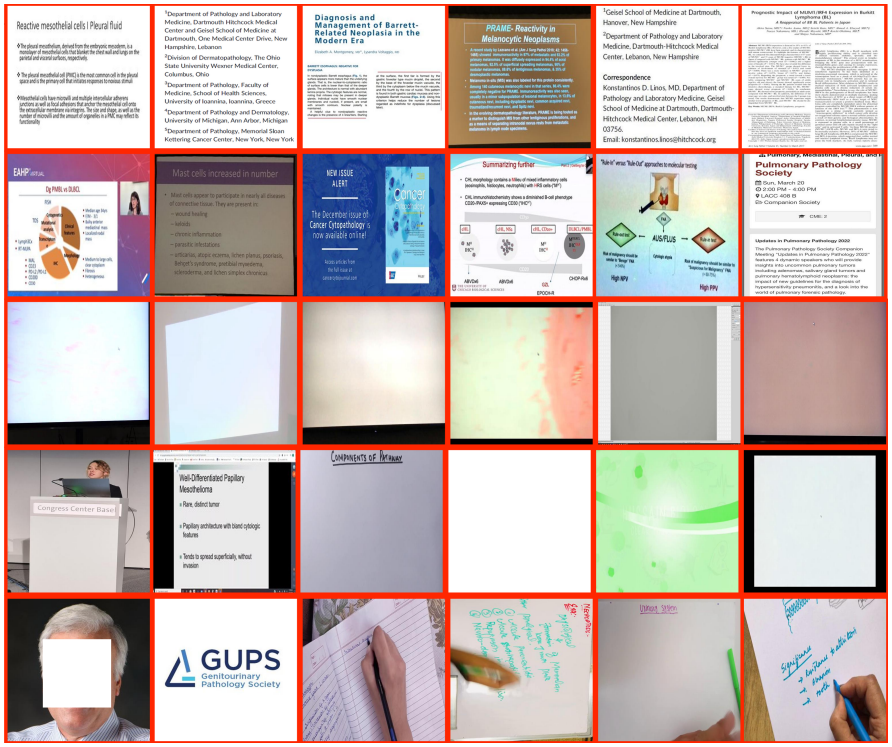


Figure 1: An example of problematic images in Quilt.

Methods introducing external knowledge often rely on external models, datasets, or a combination of both. For instance, [9] utilized high-quality external datasets, such as HQITP-350M, to train new CLIP-style models, which were subsequently employed to filter out low-quality data. Similarly, [18] introduced an image captioning model to enhance the quality of captions. Approaches like [19] and [27] targeted specific issues, such as removing images that contained embedded text.

Methods without external knowledge operate within the constraints of the existing OpenAI (OAI) model and dataset. These methods typically use existing embeddings and implement various sampling strategies or metrics to filter low-quality data. A well-known baseline in this category is CLIP score-based filtering, which removes mismatched image-text pairs. This approach has been widely adopted as a standard for data processing in works such as [10] and [22]. Another example is [2], which used embeddings from OAI models in conjunction with k-means clustering for cluster-based deduplication, effectively removing redundant data instances. [1] extended this concept further by employing cluster density-based sampling to balance the dataset, achieving more nuanced data pruning beyond mere deduplication.

A natural question arises: can these data pruning techniques for general-purpose datasets be directly applied to clean pathological datasets? Unfortunately, the answer is no. Our findings reveal that CLIP score-based methods fail to filter out non-pathological images because the CLIP scores of non-pathological image-text pairs are not necessarily lower than those of pathological pairs. Similarly, clustering-based methods and approaches incorporating image captioning models are unable to effectively exclude non-pathological images.

Therefore, it becomes evident that before applying general-domain data governance methods, we must first develop a specialized approach to filter non-pathological images from pathological image-text datasets. Only then can the broader toolkit of data pruning methods be effectively adapted for pathological datasets.

Histopathology Data Filtering In the field of pathology, the governance of image-text datasets remains relatively underexplored. However, some studies have highlighted critical issues related to image quality. For example, [5] identified eight categories of problematic images: Narrator/person, Desktop/window decorations/slide viewer, Text/logo, Arrow/annotations, Image of

insufficient quality, Additional slide overview, Additional buttons/control elements, and Multi-panel images. To address these issues, they developed a multi-label classifier based on ResNet40-D to categorize images into these groups.

Similarly, [25] manually annotated 20,000 samples as either pathological or non-pathological and trained a ConvNext-tiny model on this dataset. The trained model was then used to construct a pathology-specific dataset containing 135,000 high-quality pathology-specific images. These efforts represent steps toward improving the quality of pathological image-text datasets, yet there remains much to explore in this domain.

3 Task Formulation

The training dataset is a labled set denoted as $D = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in R^{C\times H\times W}, y_i \in \{0, 1\}$. $y_i = 1$ indicates x_i is a histopathology image, otherwise a non-pathological image. N is the number of samples. Our goal is to train a binary classifier M which can correctly classify pathological and non-pathological images.

4 Experiments

4.1 Setup

Datasets A straightforward yet effective approach to constructing a binary classification dataset for pathological images is to utilize patches extracted from Whole Slide Images (WSIs) and general-purpose datasets. It is important to note that pathological WSIs contain limited fine-grained textual information, making the development of a binary text classifier infeasible. Consequently, we focus solely on constructing binary classification datasets for pathological images. To ensure data balance during the slicing process, we uniformly selected a total of 2.01 million patches (denoted as TCGA_patches) from WSIs representing 15 cancer types within The Cancer Genome Atlas (TCGA). These cancer types include ACC, CESC, CHOL, DLBC, ESCA, GBM, KICH, KIRC, KIRP, LUSC, OV, PCPG, PRAD, SARC, and STAD. For non-pathological images, we sampled 2 million samples (denoted as CC3M_subset) from CC3M [23], a general-purpose dataset. This approach ensures a comprehensive and balanced dataset for training the binary classifier. We also construct a small dataset (denoted as TCGA_patches-tiny), which has 1/10 samples of TCGA_patches and CC3M_subset (denoted as CC3M_subset-tiny) to compare training data scales' influence.
Models We use five models mainly: Mobilenet_v2, Resnet50, Efficientnet_b3, and UNI[8] to train binary classifier. These models have different sizes of parameters to help us compare different model structures' influence on filtering performance. For Resnet50, Efficientnet_b3 and ViT-B-32, we train them from scratch. Note that UNI is a pretrained model, hence we finetune this model on the constructed dataset. Convnext [25] is a baseline for the experiment.
Metrics In the training phase of the filtering model, we employ accuracy as the evaluation metric due to the uniform sampling of both pathological and non-pathological images during the dataset construction, i.e., a balanced dataset. Additionally, we will assess the zero-shot performance of downstream tasks using vision-language models (VLM) that has been trained on the 'clean' dataset filtered by these models.
Implementation Detail In order to make the model as accurate as possible, we set different parameters to train the model. The detail information of training parameters are listed in 3.

5 Result Analysis

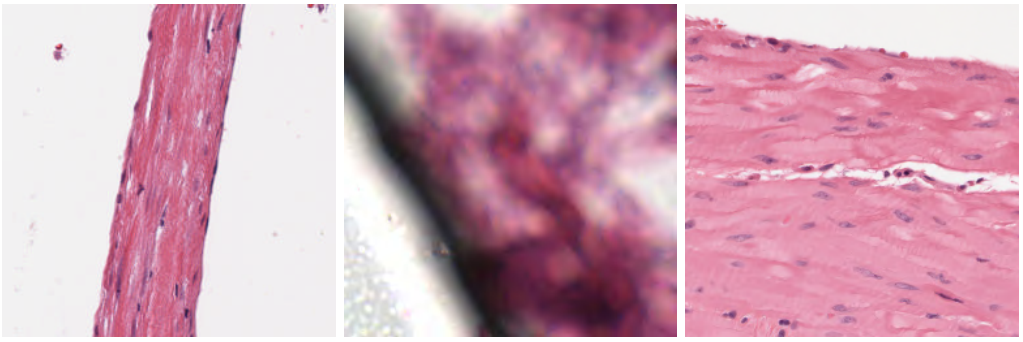
5.1 Training Result

With the exception of the vitb32 model, all the other models achieved an accuracy of more than 99.9%. Following table 1 lists the training accuracy of these models and number of misclassified samples in both TCGA_patches and CC3M_subset. Here we just show models trained on 4

million samples, since models can achieve 99.9%+ accuracy on this large dataset. For more information about the tiny dataset, refer to appendix **??**

Model	**Total**		**TCGA_patches**		**CC3M_subset**	
-	ACC	#wrong	ACC	#wrong	ACC	#wrong
Mobilenetv2	-%	-	-%	-	-%	-
Resnet50	99.95%	1701	99.99%	170	99.92%	1531
Efficientnet_b3	99.95%	1737	99.99%	113	99.92%	1624
ViT-B-32	99.84%	6105	99.91%	1796	99.78%	4309
UNI	**99.99%**	73	**99.99%**	21	**99.99%**	52
Convnext	99.59%	16253	99.36%	12948	99.83%	3305

Table 1: Results of classifications



(a) Blank samples. (b) Blurry samples. (c) Normal samples.

Misclassifications in dataset Table 1 shows that the performance of UNI is best, achieves 99.9% accuracy. The performance of Convnext is worse, espacially in classifying pathological images.
Misclassifications in CC3M_subset includes images with regular or symmetrical patterns. Those samples also prefer pink or blue color, like colors of pathological images. This shows that our classifier does learn some features of the pathological images. We find that misclassifications in TCGA_Patches could by divided into three categories: 'Blank', 'Blurry' and 'Normal'. 'Blank' refers to patches with large regions of blank, e.g., 2a. 'Blurry' means that patches have dark regions or blurry regions, e.g., 2b. 'Normal' images are those pathological images don't have aforementioned problems, e.g., 2c. For a more intuitive understanding of these three types of images, refer to the following figures. After Manual statistics, we found misclassifications caused by Resnet50 and Efficientnet_b3 models mainly consist of 'Blank' and 'Blurry'. The 'Normal' misclassifications only accounts for about 10%, which means this model lacks identification capacity for abnormal pathological images. This problem is more serious in vitb32, so there are a higher number of misclassified samples. However, in uni model, this problem not exists.

5.2 Downstream Performance

Downstream Tasks We use these filtered datasets for VLM model training and evaluate its performance on downstream tasks. Specifically, we use average weighted F1 (WF1) score of zero-shot classifications in 11 benchmarks: BACH[4], BreastPathQ[20], CRC100k[16], CRC-MSI[15], LC25000[6], MHIST[26], Pannuke[11], SICAP[24], SkinCancer[17], TCGA-TILs[3], Wsss4luad[12].
Training strategy We finetuned on pretrained models use filtered datasets. In this setup, we applied pretrained weights from UNI [8] and BioMedBERT [7] models for fine-tuning the visual and text encoders, respectively, tailored to the medical domain.
Filtering strategy We introduced a threshold denoted as ε to differentiate filtering samples under a same filtering model. A lower value of ε indicates a higher level of uncertainty regarding the non-pathological nature of a sample. We evaluated the model using three different thresholds: $\varepsilon = 0.1$, $\varepsilon = 0.5$, and $\varepsilon = 0.9$. These thresholds represent varying degrees of filtering:

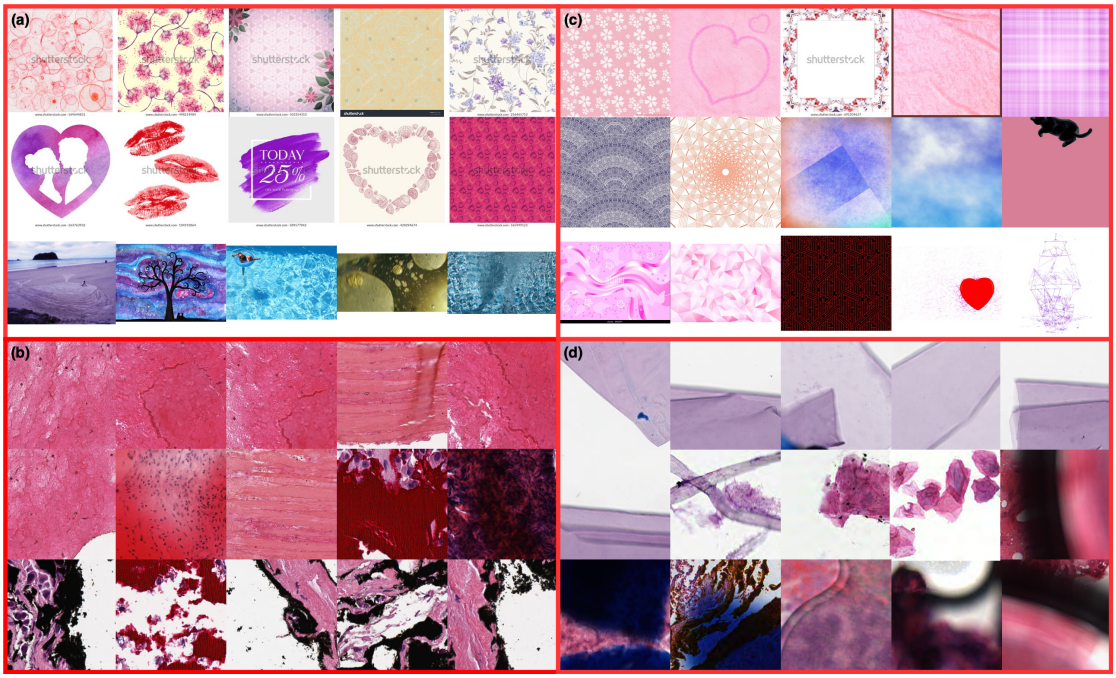
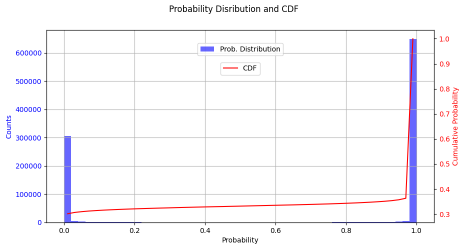
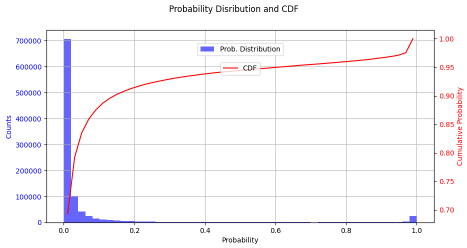


Figure 3: An example of misclassifications of Convnext. (a) shows some examples are misclassified by convnext as pathological image in CC3M_subset. (b) shows some examples are misclassified by convnext as non-pathological image in TCGA_patches. (c) shows some examples are misclassified by uni as pathological image in CC3M_subset. (d) shows some examples are misclassified by uni as non-pathological image in TCGA_patches.

$\varepsilon = 0.1$ means samples are considered highly pathological. $\varepsilon = 0.9$ means only the most obvious non-pathological samples are filtered out, leaving some ambiguous cases in the dataset. $\varepsilon = 0.5$ serves as a standard threshold for binary classification. The distribution of confidence produced by UNI and Convnext are shown in 4a,4b. 4a and 4b show two different prediction confidence distribution. The confidence degree of UNI model for samples tends to be polarized, that is to say, the prediction of UNI model for samples is more confident. However, convnext model and UNI model have great differences in the judgment of samples, and only a small number of non-pathological samples are filtered out. Here we have chosen the two distributions with the biggest differences for presentation. More information about other models are listed in appendix7.



(a) UNI predict confidence distribution.



(b) Convnext predic confidence distribution.

Zero-shot Performance We trained different filtering models on binary classification datasets of different sizes. The model with the smallest number of parameters is Mobilnet_v2, which contains only 2.22M parameters, and the model with the largest number is UNI, which contains 303M parameters. Then we get clean dataset by these filteirng models and finetune UNI_Biomedbert via these clean dataset. The finetuned UNI_Biomedbert were subsequently evaluated on 11 downstream tasks. The finetuning parameters included 50 epochs, a learning rate of 0.00001, and a batch size of 224 per GPU. Average WF1 score of these downstream tasks are listed in 1,

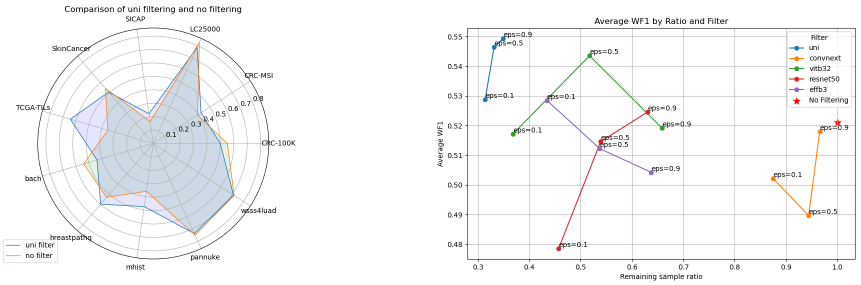
from which we can make following conclusions:

- UNI model performs best and Mobilenet_v2 performs worse when these filtering models are trained on 0.4 million samples. It is reasonable since UNI has about $\times 150$ parameters then Mobilenet_v2.
- Will a larger dataset leads to a better filtering model? The answer is YES. As we can see those large scale models have improvement on performance when they are trained on $\times 10$ data scale, i.e., 4 million samples.

The detail performance of each downstream task 5a. Note that the performance is about UNI filtering model trained on 4 million samples and $\varepsilon = 0.9$ is applied.

Finetune Model	**0.4 million samples**				**4 million samples**			
UNI_Biomedbert	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 0.9$	avg	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 0.9$	avg
Mobilenet_v2(2.22M)	47.64%	50.02%	50.35%	49.33%	-	-	-	-
Convnext(27.82M)	50.2%	49.0%	51.8%	50.33%	-	-	-	-
Resnet50(23.51M)	49.52%	49.98%	49.88%	49.79%	47.9%	51.5%	52.5%↑	50.63%(+0.84)
Efficientnet_b3(10.7M)	50.30%	51.62%	53.22%↑	51.71%	52.9%↑	51.2%	50.4%	51.50% (-0.21)
ViT-B-32(87.45M)	51.20%	**52.55%**↑	50.49%	51.41%	51.7%	54.4%↑	51.9%	52.67% (+1.36)
UNI(303M)	**55.10%**↑	52.36%↑	**54.62%**↑	54.02%	**52.9%**↑	**54.6%**↑	**54.9%**↑	54.13%(+0.09)
No filtering	52.1%							

Table 2: Filtering Model of different structure is trained on different scales of datasets, the performance of the data filtered by different ε values on the downstream task



(a) Radar chart of downstream tasks' performance

(b) Different filtering models' influence on downstream tasks

5.3 In-Depth Analysis

Confidence reflects the "pathological degree" In order to further explore the classification ability of the model for pathological images, we analyzed the cases of misclassified non-pathological images in TCGA Patches. What's more, we find that the confidence of predictions could reflect its "pathological degree". We also show cases of images with different confidence in the prediction of the model. In order to increase the flexibility of pathological image filtering, we studied the different prediction confidence corresponding to images in the pathological binary classification model. The following confidence refers to the prediction confidence of an image as non-pathological, which means an image with higher confidence is more likely to be non-pathological.

6 Conclusion

We explore a simple but effective and accurate method to filter non-pathological images from pathological datasets. The experimental results show that the patches extracted from WSIs and the images from the general dataset can be effectively used to train a classifier for distinguishing

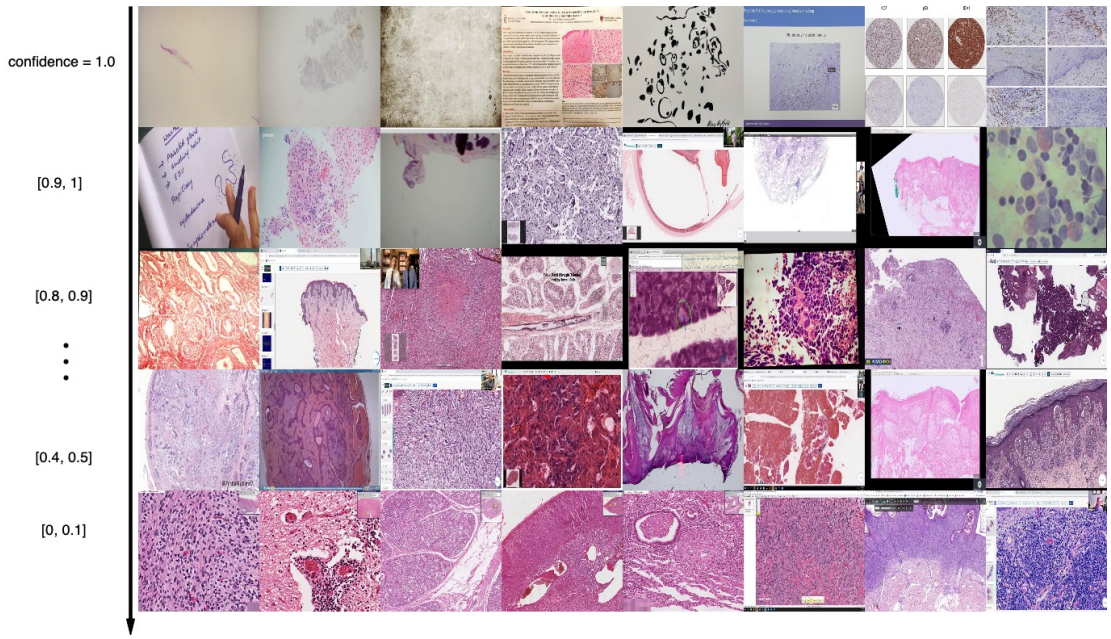


Figure 6: Filtered images in Quilt-1M change as confidence decreases

pathological images from non-pathological images. The strength of this method is that the dataset is easy to build and the classification accuracy is high. However, the disadvantage of this approach is also obvious: it requires a lot of pathological patches to train. There is stil room for exploration of methods that do not require training.

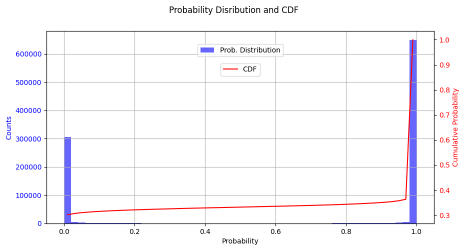
A Appendix

A.1 training setting

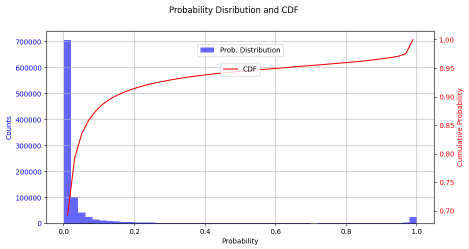
Model	Learning Rate	Batch Size	Epochs
Resnet50	0.0001	256	1
efficientnet_b3	0.0001	256	2
ViT-B-32	0.0001	256	3
UNI	0.0001	128	1

Table 3: Parameters of Model Training

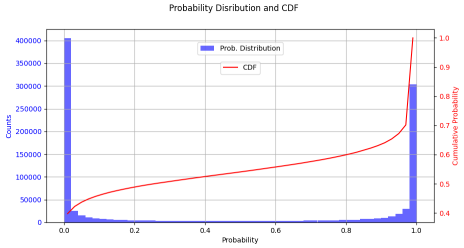
confidence distribution



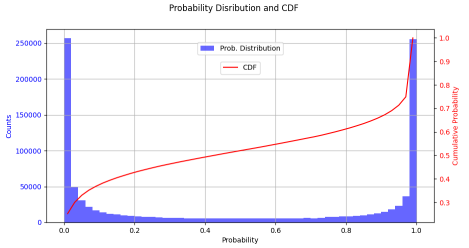
(a) uni_quilt_pred - prob. distribution



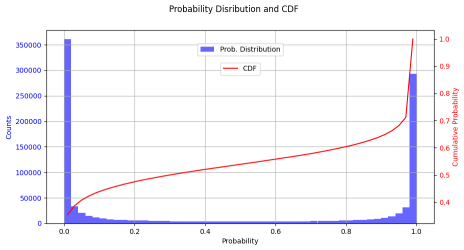
(b) convnext_quilt_pred - prob. distribution



(c) resnet50_quilt_pred - prob. distribution



(d) vitb32_quilt_pred - prob. distribution



(e) effb3_quilt_pred - prob. distribution

Figure 7: Different filtering models prediction prob. distributions

References

[1] Amro Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. *arXiv preprint arXiv:2401.04578*, 2024.

[2] Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and AriS Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*.

[3] Shahira Abousamra, Rajarsi Gupta, Le Hou, Rebecca Batiste, Tianhao Zhao, Anand Shankar, Arvind Rao, Chao Chen, Dimitris Samaras, Tahsin Kurc, and Joel Saltz. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. *Frontiers in Oncology*, 11, 2022.

[4] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[5] Marc Aubreville, Jonathan Ganz, Jonas Ammeling, Christopher Kaltenecker, and Christof Bertram. Model-based cleaning of the quilt-1m pathology dataset for text-conditional image synthesis. In *Medical Imaging with Deep Learning*.

[6] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[7] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th international conference on computational linguistics*, pages 669–679, 2020.

[8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[9] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[10] SamirYitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, PangWei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. Apr 2023.

[11] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019.

[12] Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, et al. Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022.

[13] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

[14] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024.

[15] Jakob Nikolas Kather. Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples, February 2019.

[16] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

[17] Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzelski, Jörg Kriegsmann, Charlotte Janßen, Rolf Rüdinger Meliß, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12, 2022.

[18] Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S Morcos. Sieve: Multimodal dataset pruning using image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22423–22432, 2024.

[19] Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. T-mars: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*, 2023.

[20] Nicholas A. Petrick, Shazia Akbar, Kenny H. H. Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A. Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne L. L. Martel, and for the BreastPathQ Challenge Group. SPIE-AAPM-NCI BreastPathQ Challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging*, 8(3):034501, 2021.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[22] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, 2018.

[24] Julio Silva-Rodriguez, Adrián Colomer, Jose Dolz, and Valery Naranjo. Self-learning for weakly supervised gleason grading of local patterns. *IEEE journal of biomedical and health informatics*, 25(8):3094–3104, 2021.

[25] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao, Honglin Li, Yunlong Zhang, Ruojia Zhao, et al. Pathasst: Redefining pathology through generative foundation ai assistant for pathology. *arXiv preprint arXiv:2305.15072*, 2, 2023.

[26] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.

[27] Haichao Yu, Yu Tian, Sateesh Kumar, Linjie Yang, and Heng Wang. The devil is in the details: A deep dive into the rabbit hole of data filtering. *arXiv preprint arXiv:2309.15954*, 2023.