EXPLORING FOUNDATION MODELS FINE-TUNING FOR CYTOLOGY CLASSIFICATION

Manon Dausort[⋆,1] *Tiffanie Godelaine*[⋆,1] *Maxime Zanella*[1,2] *Karim El Khoury*[1]
Isabelle Salmon[3] *Benoît Macq*[1]

[1] ICTEAM, Université Catholique de Louvain, Belgium
[2] ILIA, Université de Mons, Belgium
[3] CMMI, Université Libre de Bruxelles, Belgium

ABSTRACT

Cytology slides are essential tools in diagnosing and staging cancer, but their analysis is time-consuming and costly. Foundation models have shown great potential to assist in these tasks. In this paper, we explore how existing foundation models can be applied to cytological classification. More particularly, we focus on low-rank adaptation, a parameter-efficient fine-tuning method suited to few-shot learning. We evaluated five foundation models across four cytological classification datasets. Our results demonstrate that fine-tuning the pre-trained backbones with LoRA significantly improves model performance compared to fine-tuning only the classifier head, achieving state-of-the-art results on both simple and complex classification tasks while requiring fewer data samples. Our source code is available on GitHub https://github.com/mdausort/Cytology-fine-tuning.

***Index Terms*—** Cytology classification, Parameter-efficient Fine-tuning, Low-rank Adaptation, Transfer Learning, Few-shot Learning

1. INTRODUCTION

Cytology slides are vital for diagnosing and staging cancer, offering detailed views of abnormal cells to guide treatment decisions [1]. Analyzing these slides is labor- intensive and costly, leading to delays in the reporting process, making automation essential to improve both classification efficiency and accuracy [2].

An effective approach to automate classification involves using deep learning models. However, the performance of these models is highly dependent on the quantity and diversity of data [3]. Since handling large-scale cytology images demands high computational resources, a patch-based method is often chosen to divide gigapixel slides into thousands of smaller patches suitable for model training [4]. Annotating thousands of patches is cumbersome and challenging, as it must be performed by trained medical practitioners and requires a considerable amount of time [2]. As a result, the community is exploring effective ways to utilize limited labeled data, an area of research known as few-shot learning [5]

In parallel, large-scale models often referred to as foundation models (FMs) are being developed through extensive pre-training on large datasets [6, 7] and are valued for their robust generalization capabilities, especially in few-shot regimes. They have the potential to address data scarcity in more specific downstream tasks. Hence, the medical imaging community has been gathering large datasets to exploit the potential of foundational models, leading to medical imaging-oriented FMs [8–12] and paving the way for further research extensions in the field [13, 14]. In the context of digital pathology, histology has garnered considerable attention leading to the introduction of several FMs specifically designed for histological analysis [9–12]. For instance, UNI [12] was pre-trained on an extensive dataset comprising over 100 million images across 20 tissue types, totaling 77 TB of data. In contrast, cytology remains underexplored, largely due to the scarcity of large publicly available datasets. In comparison, the most extensive cytology dataset available to the public is composed of 40,229 images yet it only covers a single pathology [15]. Due to the extensive data, time, and energy consumption required for training those FMs, we are encouraged to consider fine-tuning existing FMs as an efficient solution for cytology classification.

Traditionally, fine-tuning these models relies on computationally intensive methods. However, recent parameter-efficient fine-tuning approaches reduce the number of trainable parameters, enhancing generalization with minimal labeled data [16, 17]. Among these methods, Low Rank Adaptation (LoRA) [17] allows model training without increasing the number of parameters at inference and has shown strong potential in low-data scenarios across various medical tasks and modalities [18]. Motivated by the potential of LoRA demonstrated in previous studies, we aim to extend its application to cytology classification.

In this work, we investigate the applicability of existing FMs for cytology tasks through three experiments. First, we assess whether domain-specific models offer greater adapt-

⋆The authors have contributed equally to this work.

ability than general-purpose ones and explore the potential for transferring knowledge from histology to cytology. Then, we examine the benefits of using LoRA fine-tuning in few-shot scenarios. Finally, we fine-tune the most promising vision encoder with increasing amount of data to achieve state-of-the-art performance. We evaluate five FMs pre-trained on natural, biomedical and histopathological images on four cytology classification benchmark datasets and **observe that**:

- Histology-specific models show superior adaptability as feature extractors for cytology classification.
- General models achieve better generalization when their weights are fine-tuned with LoRA in few-shot settings.
- CLIP's vision encoder fine-tuned with LoRA achieves state-of-the-art performance on a challenging dataset while utilizing only 70% of the dataset and requiring far fewer trainable parameters than the current state-of-the-art model.

2. RELATED WORK

2.1. Cytology classification

Many cytology classification studies use pre-trained CNN models as feature extractors, typically fine-tuning only the classification layer added on top of the backbone [19, 20]. For example, UÇA *et al.* [21] compared various CNN architectures, finding ResNet50 effective for body cavity cytologies. Yaman and Tuncer [22] employed DarkNet19 to extract features from cervix cytological images, combined with traditional machine learning techniques. While CNNs remain prevalent, transformer-based models have recently gained attention for their ability to capture complex feature relationships [19, 20]. Cai *et al.* [15] introduced HierSwin, a model based on Swin Transformer architecture that leverages hierarchical information, marking a shift toward more advanced transformer architectures in cytology.

2.2. Foundation models

Certainly one of the most popular FM is the vision transformer introduced by Google [7]. Beyond vision, FMs have been developed for multiple modalities, including vision and language. Vision-language models, such as CLIP [6], learn to align visual and textual embedding in a common embedding space using large-scale image-text datasets. Following the introduction of the first FMs, increasingly specialized datasets have been developed, leading to the emergence of domain-specific FMs. Examples include the vision-language models BiomedCLIP [8] and QUILT [10], focusing on biomedical and histopathological applications, respectively, as well as UNI [12], a vision transformer also designed for histopathology. To the best of our knowledge, no FM has yet been introduced specifically for cytology.

2.3. Parameter-efficient fine-tuning

As fine-tuning these models requires extensive datasets and computational resources, parameter-efficient fine-tuning methods have gained popularity for training only a small subset of parameters, as highlighted in the review by Han *et al.* [23]. Among these methods, Low Rank Adaptation (LoRA) [17] updates models weight using low-rank matrices. Dutt *et al.* [18] evaluated various parameter-efficient fine-tuning methods on transformer-based models on different medical tasks and modalities. Their findings show that LoRA consistently outperforms full fine-tuning, especially in low-data scenarios, establishing it as the most effective parameter-efficient fine-tuning method across all tested datasets. Building on this, Zanella & Ben Ayed [24] applied LoRA specifically to CLIP, demonstrating its ability to achieve strong generalization further supporting LoRA's applicability across diverse domains.

3. EXPERIMENTAL SETUP

3.1. Fine-tuning methods

To adapt FMs to cytology classification, we compare two fine-tuning strategies.

Linear classifier. A common fine-tuning approach is to leverage a pre-trained model θ to extract key features $z = f_\theta(x)$ from an input x and to add a linear classifier on top. During training, only the weights of the linear layer are updated to tailor the model for the specific task, as described by the following equation for the forward pass:

$$y = \mathbf{W}z \tag{1}$$

where $z \in \mathbb{R}^n$ and $\mathbf{W} \in \mathbb{R}^{c \times n}$, with n the number of extracted features by the model and c the number of classes.

LoRA. LoRA [17] is a parameter efficient fine-tuning method describing the update of pre-trained weights of model with low-rank matrices. Given the initial model weight $\mathbf{W} \in \mathbb{R}^{m \times n}$, the weight update $\Delta\mathbf{W} \in \mathbb{R}^{m \times n}$ is modeled as the multiplication of two matrices of small rank r, $\mathbf{A} \in \mathbb{R}^{r \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times r}$. This gives rise to a modification of the forward pass:

$$h = \mathbf{W}x + \gamma\Delta\mathbf{W}x = \mathbf{W}x + \gamma\mathbf{B}\mathbf{A}x \tag{2}$$

with h a hidden state from an intermediate layer, x the input of this layer and γ a scaling factor. Matrix $\mathbf{A}$ is randomly initialized using Kaiming initialization, whereas matrix $\mathbf{B}$ is set to all zeros. The entries of these low-rank matrices serve as the trainable parameters, while the remaining weight matrices remain frozen during training, reducing the number of parameters to be trained. During inference, this number is unchanged, as weight update $\Delta\mathbf{W}$ is added to the initial one. For instance, this can be applied to all or a subset of the attention matrices of the model (query, key, value and output) [24].

3.2. Datasets

We evaluate the classification performance on four publicly-available cytological datasets.

The **Body Cavity Fluid Cytology** [25] (BCFC) dataset consists of isolated cell images from body fluid effusions. These cells are classified into two categories: benign or malignant. The **Mendeley LBC Cervical Cancer** [26] (MLCC) dataset is composed of pap smears images of liquid based cytology, representing the four sub-categories of cervical lesions (malignant and pre-malignant) of the Bethesda System standard. **SIPaKMeD** [27] is a cervical cancer dataset, containing manually cropped images of isolated cells coming from pap smear slides. They are divided into five classes of cells. **HiCervix** [15] is a multi-center dataset of cervical cells extracted from whole slide images, resulting in 40,229 images. This makes HiCervix the largest publicly-available cervical cytology dataset. The classes are structured within a three-level hierarchical tree to capture detailed subtype information. In an effort to align with the state-of-the-art, we choose to focus on the third level of annotation, which consists of 25 classes.

3.3. Models

We examine five FMs pre-trained on natural and medical images. Among these, three are vision-language models; note that in this study, we exclusively used their visual encoders.

CLIP [6], a vision-language model trained on a dataset of 400 million image-text pairs sourced from the Internet, provides a strong foundation for general visual-language tasks. **ViT** [7], introduced by Google, is a vision transformer pre-trained on ImageNet-21k and fine-tuned on ImageNet 2012, making it highly effective for feature extraction in downstream tasks. **BiomedCLIP** [8] is a vision-language model pre-trained on PMC-15M, a dataset with a wide range of medical image modalities, making it particularly relevant for medical applications. **QUILT** [10], is a CLIP model fine-tuned on Quilt-1M, the largest vision-language dataset specifically focused on the histopathology. **UNI** [12] is a vision transformer more specific to medical as it is pre-trained on more than 100 millions histopathological images. It should be noted that, unlike previous models which use a ViT-B/16 backbone, UNI employs a ViT-L/16, composed of four times more parameters.

4. EXPERIMENTS

Results are averaged over 3 seeds. Top-1 accuracy is evaluated on the validation set to determine the best learning rate for each model-shots-dataset combination, and final reported performances are measured on the test set of each dataset. The batch size is set to 32.

Table 1. Mean accuracy of fine-tuned classifiers, evaluated on five models and four datasets.

	Models				
Datasets	CLIP	QUILT	BiomedCLIP	UNI	ViT
BCFC	94.74 ±0	**100±0**	96.17±0	99.04±0	98.72±0.28
MLCC	93.95 ±0.52	97.95±0	95.1±0.52	**99.2±0.4**	96.8±0.2
SIPaKMed	90.5 ±0.16	92±0.57	87.62±0.33	**93.33±0.16**	89.71±0.49
HiCervix	55.2±0.2	54.64±0.08	46.39±0.1	**64.07±0.25**	57.64 ±0.14
Average	83.59	86.15	81.32	**88.91**	85.72

4.1. Experiment 1: Linear classifier

Experiment details. To assess the effectiveness of existing FMs in extracting discriminative features for cytological classification, we use the aforementioned models as feature extractors, keeping the backbone frozen and fine-tuning only a classification head (Eq. (1)).

Results. The results are summarized in Table 1. On average, UNI outperforms the other models, ranking highest on three out of four datasets, while QUILT achieves better results on the first dataset. It's worth noting that UNI utilizes a ViT-L/16 backbone, resulting in four times more parameters compared to the other models. Excluding UNI, the second-best model on average across the four datasets is QUILT. The results demonstrate the effectiveness of histopathology-pre-trained FMs to extract discriminant features for cytology tasks. It should be noted that, for the first two datasets [25, 26], UNI with a fine-tuned classifier achieves accuracies consistent with those reported in the literature.

4.2. Experiment 2: LoRA few-shot adaptation

Experiment details. The objective is to determine the gain of fine-tuning FMs beyond the classification head. Due to the large number of parameters in these models, we apply LoRA (Eq. (2)) to the query and value matrices of each attention block in the visual encoder. Based on the work of Zanella and Ben Ayed [24], the rank of the LoRA matrices is set to 2. The model is trained in a few-shot setting with shots per class ranging from $1, 2, 4, 8, 16$, to 50. For instance, 2 shots of the SIPaKMeD dataset results a total of 10 examples (two examples per class).

Results. The results can be seen in Fig.1. CLIP fine-tuned with LoRA outperforms the linear classifier with only one or two samples per class on the first three datasets. For the HiCervix dataset, 50 shots are needed to achieve similar results, though this still represents only 4.4% of the total data. Models pre-trained on medical or histopathological images require more data to reach comparable performance to those trained on more diverse datasets, emphasizing the strong generalization of models with broader pre-training when samples are limited. Overall, fine-tuning the entire backbone con-

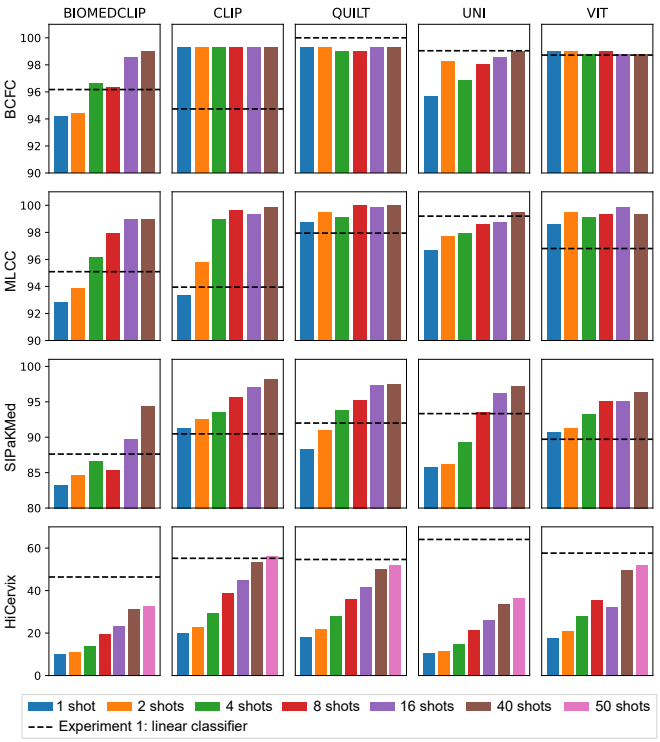

Fig. 1. Mean accuracy of fine-tuned foundation models, evaluated on four datasets in a few-shot setting with a varying numbers of shots. The black horizontal line reports the mean accuracy from Experiment 1.

sistently yields higher accuracy than fine-tuning the classifier head alone, highlighting the advantage of backbone fine-tuning in data-limited scenarios.

4.3. Experiment 3: Pushing model fine-tuning limits

Experiment details. Our goal is to determine whether the most generalizable model can achieve state-of-the-art performance using a smaller dataset proportion and fewer trainable parameters compared to the current state-of-the-art model. We focus on the HiCervix dataset, which is particularly large and presents a complex classification task due to its high number of classes. Based on Experiment 2, we select CLIP as the model with the best generalization capabilities. To further improve performance, we use a ViT-L/14 backbone and apply LoRA to the query, value, key, and output matrices with a rank of 16. It is important to note that, while ViT-L/14 is larger than the HierSwin model, the use of LoRA allows training only a portion of the parameters, reducing the number of trainable parameters to 3.1 million, which is 62 times fewer than current state-of-the-art model. Mean accuracy was evaluated across different dataset proportions, ranging from 5% to 100%, and compared to HierSwin's state-of-the-art performance [15].

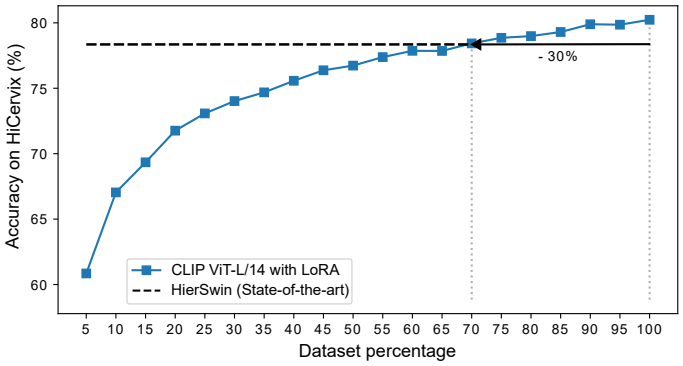

Fig. 2. Mean accuracy of CLIP fine-tuned with LoRA and a ViT-L/14 backbone on different percentage of the HiCervix dataset. The black horizontal line represents the state-of-the-art accuracy (HierSwin).

Results. The results are shown in Fig.2. We observe that the model fine-tuned on 70% of the HiCervix dataset achieves state-of-the-art performance, outperforming HierSwin in terms of both parameter efficiency and data usage. Specifically, this setup allows for high accuracy while updating significantly fewer parameters and utilizing fewer samples than required by HierSwin. When trained with the full dataset (100%), the model's accuracy further increases, reaching 80.23%, suggesting that collecting more data could further improve model's performance. These findings highlight the model's ability to achieve competitive accuracy with a reduced training load.

5. CONCLUSION

This study investigates the potential of foundation models for cytology. We show that fine-tuning with LoRA significantly improves performance over classifier-only fine-tuning, with particularly strong results in few-shot settings where labeled data is limited, and further gains when more data is available. Our findings suggest that while histology and cytology share visual features, models pre-trained on histology images do not necessarily generalize better for cytology. This is likely due to the absence of tissue structures in cytology, underscoring the need for cytology-specific adaptations in foundation models.

Future work could investigate the benefits of incorporating textual information through vision-language models, potentially enhancing model performance. Overall, we demonstrate that fine-tuning existing foundation models is deployable in cytology classification while remaining both practical and resource-efficient. Such adaptations could enhance the accuracy and accessibility of AI-driven cytology classification, supporting its growing role as a critical diagnostic tool.

6. COMPLIANCE WITH ETHICAL STANDARDS

This is a study for which no ethical approval was required.

7. ACKNOWLEDGMENTS

M. Dausort and T. Godelaine are funded by the MedReSyst project, supported by FEDER and the Walloon Region. M. Zanella is funded by the Walloon region under grant No. 2010235 (ARIAC by DIGITALWALLONIA4.AI). Computational resources were made available on the Lucia infrastructure (Walloon Region grant n°1910247).

8. REFERENCES

[1] Patricia A Shaw, "The history of cervical screening i: the pap. test," *Journal SOGC*, vol. 22, no. 2, pp. 110–114, 2000.

[2] Hao Jiang, Yanning Zhou, et al., "Deep learning for computational cytology: A survey," *Medical Image Analysis*, vol. 84, pp. 102691, 2023.

[3] Zhiqiang Gong, Ping Zhong, et al., "Diversity in machine learning," *Ieee Access*, vol. 7, pp. 64323–64350, 2019.

[4] Ling Zhang, Le Lu, et al., "Deeppap: deep convolutional networks for cervical cell classification," *IEEE journal of biomedical and health informatics*, vol. 21, no. 6, pp. 1633–1643, 2017.

[5] Yisheng Song, Ting Wang, et al., "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–40, 2023.

[6] Alec Radford, Jong Wook Kim, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] Bichen Wu, Chenfeng Xu, et al., "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[8] Sheng Zhang, Yanbo Xu, et al., "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[9] Zhi Huang, Federico Bianchi, et al., "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[10] Wisdom Ikezogwo, Saygin Seyfioglu, et al., "Quilt-1m: One million image-text pairs for histopathology," *Advances in neural information processing systems*, vol. 36, 2024.

[11] Ming Y Lu, Bowen Chen, et al., "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863–874, 2024.

[12] Richard J Chen, Tong Ding, et al., "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[13] Maxime Zanella, Fereshteh Shakeri, et al., "Boosting vision-language models for histopathology classification: Predict all at once," in *International Workshop on Foundation Models for General Medical AI*. Springer, 2024, pp. 153–162.

[14] Ruiwen Ding, James Hall, et al., "Improving mitosis detection on histopathology images using large vision-language models," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[15] De Cai, Jie Chen, et al., "Hicervix: An extensive hierarchical dataset and benchmark for cervical cytology classification," *IEEE Transactions on Medical Imaging*, 2024.

[16] Elad Ben Zaken, Shauli Ravfogel, et al., "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199*, 2021.

[17] Edward J Hu, Yelong Shen, et al., "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[18] Raman Dutt, Linus Ericsson, et al., "Parameter-efficient fine-tuning for medical image analysis: The missed opportunity," *arXiv preprint arXiv:2305.08252*, 2023.

[19] Hao Jiang, Yanning Zhou, et al., "Deep learning for computational cytology: A survey," *Medical Image Analysis*, vol. 84, pp. 102691, 2023.

[20] Peng Jiang, Xuekong Li, et al., "A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis," *Artificial Intelligence Review*, vol. 56, no. Suppl 2, pp. 2687–2758, 2023.

[21] Murat UÇA, Buket Kaya, et al., "Comparison of deep learning models for body cavity fluid cytology images classification," in *2022 International Conference on data analytics for business and industry (ICDABI)*. IEEE, 2022, pp. 151–155.

[22] Orhan Yaman and Turker Tuncer, "Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images," *Biomedical Signal Processing and Control*, vol. 73, pp. 103428, 2022.

[23] Z. Han, C. Gao, J. Liu, et al., "Parameter-efficient fine-tuning for large models: A comprehensive survey," 2024.

[24] Maxime Zanella and Ismail Ben Ayed, "Low-rank few-shot adaptation of vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1593–1603.

[25] P. Sanyal, "Body cavity fluid cytology images," `https://www.kaggle.com/datasets/cmacus/body-cavity-fluid-cytology-images`.

[26] Elima Hussain, Lipi B Mahanta, et al., "Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions," *Data in brief*, vol. 30, pp. 105589, 2020.

[27] Marina E Plissiti, Panagiotis Dimitrakopoulos, et al., "Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 3144–3148.