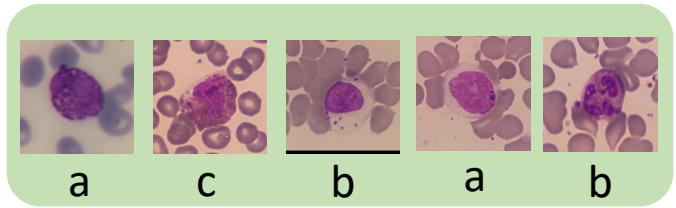
# Additional Knowledge Encoding




Downstream training set:
Images + Classification Labels

Model

Model

Fine-Tuning without additional information

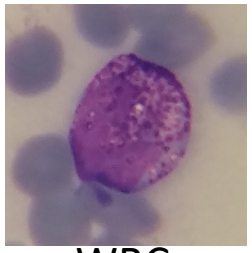
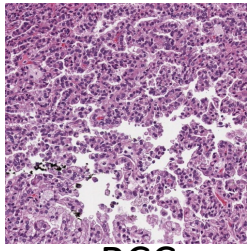
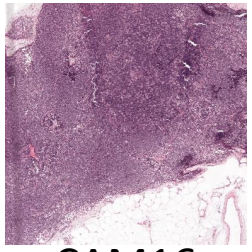
Fine-Tuning with additional information

Model

Model

This one should be better right?

# Dataset

| Scale | Dataset name | |
|---|---|---|
| S |  | WBC |
| M |  | pRCC |
| L |  | CAM16 |

Those dataset come with labelled images by senior pathologists.
The classification task are supported with (classification labels 100% & segmentation masks 10%).

The Raabin-WBC dataset [1] is abbreviated as WBC. This S scale dataset is composed of microscopic images of white blood cells, with each image containing only one or two stained white blood cells. For the downstream classification tasks, this dataset comprises 301 basophil, 1066 eosinophil, 3461 lymphocyte, 795 monocyte, and 8891 neutrophil images.

The pRCC dataset [2] consists of Papillary Renal Cell Carcinoma subtyping images, selected and cropped by pathologists from the TCGA-KIRP dataset. This dataset comprises 870 type 1 ROIs and 547 type 2 ROIs, with each image meeting the M scale dataset criteria.

The Camelyon16 dataset [3] is abbreviated as CAM16 in this paper. This WSI dataset is derived from the Cancer Metastases in Lymph Nodes challenge. In each WSI, we select 5 to 10 ROIs with dimensions of 8000*8000 (2560*2560 under the CPIA standard) to meet the average L scale dataset criteria. Our final CAM16 dataset comprises 540 tumor and 541 normal images.

# Machine Learning Task:

Given three data sets WBC, pRCC, Camelyon16. Perform classification on the WBC dataset with classes: basophil, eosinophil, lymphocyte, monocyte and neutrophil. Use the pRCC and Camelyon16 dataset for pre-training. A fraction of the training set is given with annotation mask.

Main task: You need to develop an algorithm take advantage of additional information to improve your model's performance.
Subtask1: Use 100% of WBC training set for training. Perform training with and without additional information.
Subtask2: Use 50% of WBC training set for training. Perform training with and without additional information.
Subtask3: Use 10% of WBC training set for training. Perform training with and without additional information.
Subtask4: Use 1% of WBC training set for training. Perform training with and without additional information.

We will segregate the different training set and put them in different folder.

# Report and grading:

You need to submit a report, trained model and code. Report should consist of the following sections with the following grad distribution (with a total of 15%). max 6 pages with figures

1. [4%] Background of the project, background of the data set, the ML task and its impact to the world.
2. [4%] Description of the method, we look for new ideas and creativity
3. [3%] Results with plots showing that your code is working
4. [4%] Conclusion and describe your scientific discoveries

# Ref


Kouzehkanan, Z.M., Saghari, S., Tavakoli, S., Rostami, P., Abaszadeh, M., Mirzadeh, F., Satlsar, E.S., Gheidishahran, M., Gorgi, F., Mohammadi, S. and Hosseini, R., 2022. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. Scientific reports, 12(1), p.1123.

Gao, Z., Hong, B., Zhang, X., Li, Y., Jia, C., Wu, J., Wang, C., Meng, D. and Li, C., 2021. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24 (pp. 299-308). Springer International Publishing.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R. and Manson, Q.F., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience, 7(6), p.giy065.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R., 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

Chen, X., Xie, S. and He, K., “An Empirical Study of Training Self-Supervised Vision Transformers. arXiv preprint arXiv:2104.02057

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).

Zhang, T., Yan, Z., Li, C., Ying, N., Lei, Y., Feng, Y., Zhao, Y. and Zhang, G., 2023. CellMix: A General Instance Relationship based Method for Data Augmentation Towards Pathology Image Analysis. *arXiv preprint arXiv:2301.11513*.

Zhang, T., Feng, Y., Feng, Y., Zhao, Y., Lei, Y., Ying, N., Yan, Z., He, Y. and Zhang, G., 2022. Shuffle Instances-based Vision Transformer for Pancreatic Cancer ROSE Image Classification. *arXiv preprint arXiv:2208.06833*.