# CS5242 Project

## Introduction

Hematologic cell identification, a critical aspect of medical diagnosis and treatment, presents a compelling challenge in the modern healthcare landscape. The ability to accurately classify various blood cell types can have a profound impact on patient care, disease monitoring, and research advancement. However, this task is complex due to the intricate morphological features of different cell types and the inherent variability within microscopic images.

The primary objective of this project is to develop a classification system capable of distinguishing between five classes of white blood cells: basophil, eosinophil, lymphocyte, monocyte, and neutrophil. To achieve this, we will adopt a two-phase approach. In the first phase, we will leverage pRCC and the labeled Camelyon16 datasets for pre-training.

The second phase of the project will focus on fine-tuning and classification using the WBC dataset. This dataset, containing microscopic images of blood cells, poses unique challenges due to the subtle differentiations required for accurate classification. Training data will come with annotation masks, providing valuable guidance for model training.

Through this project, you will have the opportunity to engage with real-world medical data, apply cutting-edge techniques in image analysis, and navigate the intricacies of transfer learning.

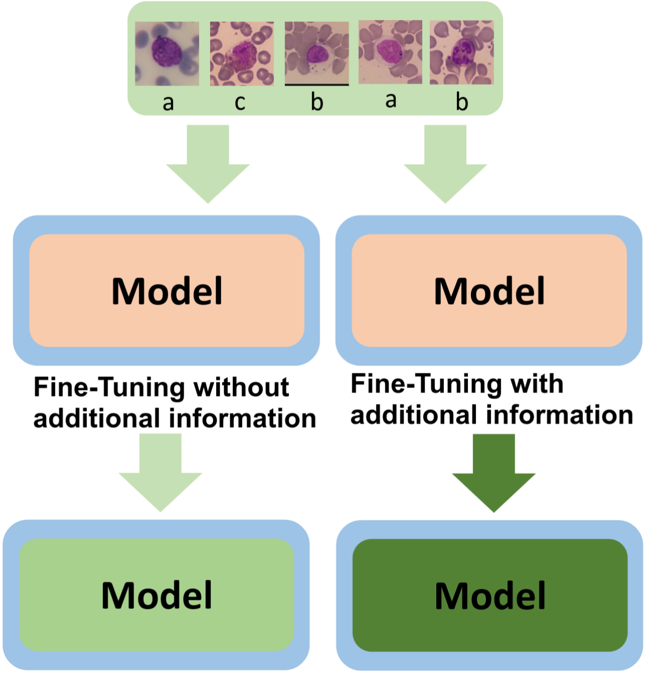
a
c
b
a
b
Model
Model
Fine-Tuning without additional information
Fine-Tuning with additional information
Model
Model

Figure 1: Additional Knowledge Encoding

## Dataset overview

*WBC Dataset*:

This dataset comprises microscopic images of different types of white blood cells, including basophils, eosinophils, lymphocytes, monocytes, and neutrophils. Each image is associated with a specific cell type, forming the basis for our classification task.

*pRCC Dataset*:

The pRCC dataset contains a diverse range of medical images, serving as a valuable source for pre-training our classification model. While not directly related to hematologic cells, the features learned from this dataset can potentially aid in recognizing patterns and improving generalization.

*Camelyon16 Dataset*:

Similarly, the Camelyon16 dataset, although not inherently tied to hematologic cell classification, contributes to the pre-training process by exposing the model to additional medical imaging data. The insights gained from this dataset can play a pivotal role in enhancing the model's ability to capture intricate details.

Table 1: Statistics of WBC_100 dataset along with its 50% segregation, 10% segregation, and 1% segregation.

| | WBC_100 | | | WBC_50 | | WBC_10 | | WBC_1 | |
|---|---|---|---|---|---|---|---|---|---|
| Class | Train | | Validation | Train | | Train | | Train | |
| | data | mask | data | data | mask | data | mask | data | mask |
| Basophil | 176 | 17 | 36 | 88 | 8 | 17 | 1 | 1 | 0 |
| Eosinophils | 618 | 61 | 126 | 309 | 30 | 61 | 6 | 6 | 0 |
| Lymphocyte | 2015 | 201 | 412 | 1007 | 100 | 201 | 20 | 20 | 2 |
| Monocyte | 466 | 46 | 95 | 233 | 23 | 46 | 4 | 4 | 0 |
| Neutrophil | 5172 | 517 | 1059 | 2586 | 258 | 517 | 51 | 51 | 5 |
| Total# | 8447 | 842 | 1728 | 4223 | 419 | 842 | 82 | 82 | 7 |

Table 2: Statistics of CAM16 and pRCC datasets.

| | CAM16 | | | | pRCC |
|---|---|---|---|---|---|
| Class | Train | | Validation | Test | Train |
| | data | mask | data | data | data |
| normal | 379 | 37 | 54 | 108 | 1419 |
| tumor | 378 | 37 | 54 | 108 | |
| Total# | 757 | 74 | 108 | 216 | 1419 |

*Dataset Statistics*:

We summarize the statistics of datasets WBC, CAM16 and pRCC in Table 1 and Table 2. In WBC_100, we provide a ratio of 5:1 data for training and validation set in each cell type. And there are three segregations (*i.e.*, WBC_50, WBC_10, WBC_10) for WBC_100, which contain 50%, 10% and 1% data of the whole set, respectively. Both WBC and CAM16 have additional mask annotation, where 10% of the samples have masks, each name corresponding to the image name. Note, pRCC and CAM16 are offered as the pre-training set, and pRCC comes without any label. The organized dataset can be found in: Link.

## Tasks

Given three datasets WBC, pRCC, and Camelyon16, the objective is to perform classification on the WBC dataset with classes:

- Basophil
- Eosinophil
- Lymphocyte
- Monocyte
- Neutrophil

Formally, each dataset $\mathcal{D}_i = \{(x, y)\}^P$ is composed of $P$ pairs of images $x \in [0, 255]^{N \times M \times 3}$ and corresponding masks $y \in \{0, 1\}^{N \times M}$, where $N$ and $M$ represent the dimensions of an image. The objective of this project is to develop a classification model $f(x, y) \in \mathbb{R}$ by leveraging the datasets pRCC and Camelyon16 for pre-training purposes (as illustrated in Figure 1). The primary focus is to investigate the impact of incorporating supplementary information on the performance of classification on WBC. Therefore, the problem will be decomposed via four different sub-tasks:

1. Use 100% of WBC training set for training. Perform training with and without additional information.
2. Use 50% of WBC training set for training. Perform training with and without additional information.
3. Use 10% of WBC training set for training. Perform training with and without additional information.
4. Use 1% of WBC training set for training. Perform training with and without additional information.

Subsample datasets are made available on Canvas.

## Submission instructions

Submit report and codes on canvas: *ProjectMatriculCode.zip*

## Project grading

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

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

Note: We will open a Kaggle competition after project deadline so that everyone can check the performance of their methods and even achieve improvement through communication. The reason for not opening it early is to prevent everyone from focusing on overfitting the test data.