CS5242 Project – Classifying White Blood Cells
Mo Yunbin e0556074

1. Background

White blood cells are an indispensable type of cells in human body acting as defensive armies for the human body. There are five basic types of WBC, namely Neutrophils, Eosinophils, Basophils, Lymphocytes, Monocytes, each type has different functions and mechanisms for protecting the human body from infections and foreign pathogens. Therefore, WBC count and differential are important indicators in diagnosis of many diseases as the elevation or reduction in the amount of certain type of WBC reflects potential abnormalities. However, manual inspection of microscopic images could be inefficient and error-prone, which brings up the needs for accurate and automated approach to identify WBC. Therefore, the goal of this project is to build a machine learning model capable of distinguishing different types of white blood cells (WBC) given their images.

With advancements in machine learning techniques, we can train powerful models for computer vision tasks in a supervised manner given relevant datasets. Three datasets of different image scale are used in this project, they are WBC dataset, pRCC dataset and the Camelyon16 (CAM16) dataset. The WBC dataset consists of microscopic images of 5 types of WBC, with each image containing one or two stained white blood cells and it is the smallest scale dataset among the three. The second dataset contains Papillary Renal Cell Carcinoma (pRCC) subtyping image, pRCC is a kind of kidney cancer and it is classified as type 1 of type 2 based on histomorphology features. This medium-scale dataset comes with no label and both type 1 and type 2 images are present. Lastly, the CAM16 dataset contains images of cancer metastases in lymph node, and each image is labelled as either 'normal' or 'tumour'. These images are of high-resolution and are consider large-scale. Moreover, for WBC and CAM16, 10% of the training data comes with masks that segments the WBC or tumour from the background.

The machine learning task in this project is to train an image classification model such that given an image of a WBC, the model can assign it a label $\in$ {Neutrophils, Eosinophils, Basophils, Lymphocytes, Monocytes}. This project also explored the possibility and effectiveness of pre-training on similar datasets to boost the performance of the classifier.

Solving this ML task with high classification accuracy will greatly benefit the treatment workflow such as early disease detection and diagnosis. Moreover, it could make diagnosis services more accessible, reduce cost of manual analysis and provide meaningful resources for educating health professionals. The automated classification can also serve as an input to further medical research such as drug development.

2. Method

This project adopted a two-stage approach where the first stage is to pre-train on pRCC and CAM16 datasets and the second stage is to perform classification on white blood cells by fine-tuning on the WBC datasets. The project also explored the effectiveness of pre-training on similar yet not directly related microscopic medical image data and impact of the size of the datasets on the classification performance.

2.1 No additional information

To effectively extract intricate features of different cells, a pre-trained ResNet-18 was used. It is shown by many studies that ResNet is effective in resolving the network degradation problem

in deep neural networks with identity mapping and it is good at learning meaningful patterns from images, which is a suitable backbone for fine-tuning for downstream tasks.

As medical images are quite different from the ImageNet images that ResNet was trained on, training the network from scratch might cause the pre-trained model to lose its generalization ability learned from ImageNet. Therefore, to fine-tune it on the WBC datasets, only the last two 2-layer convolutional blocks are opened for weight updates (as shown in Figure 1.) to learn patterns specific to white blood cells while the rest layers are frozen during training as they contain more general features such as edges and textures that we want to keep. Moreover, the number of output features of the last fully connected layers is changed from 1000 (number of classes in ImageNet) to 5 (number of classes in WBC) to suit our task.

To enhance the model's generalization ability, common data augmentations techniques were applied, including flipping and rotation and standard data normalization strategy was used to help the network learn better and faster.

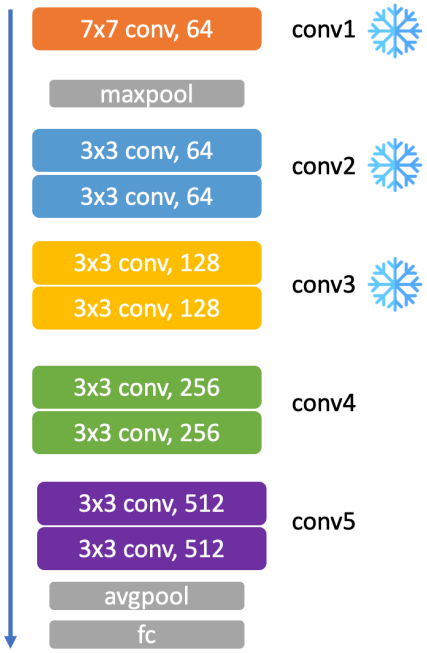

Figure 1.

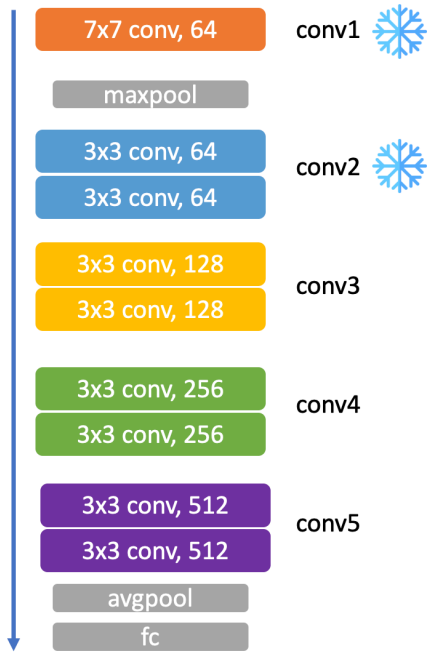

Figure 2.

2.2 Adding additional information

Although pRCC and CAM16 datasets are not directly related to white blood cells, they provide additional microscopic medical images which contain helpful underlying patterns common to the WBC dataset.

The CAM16 dataset contains images of two classes, namely "tumor" and "normal", we can train a binary classifier so that meaningful high-level features can be obtained when the model learns how to classify different images. For this task, a pretrained ResNet-18 was used. However, since medical images are unnatural, unlike common objects frequently seen in the real world, some of the general features are not present in medical images, therefore only the first convolutional layer and the first 2-layer convolutional block were frozen during training (as shown in Figure 2.). In this way, the network retains basic image recognition probability while adapts to patterns in medical images. In addition, the input images are masked with the

provide mask image where applicable, in order to make it easier for the model to focus on important areas. Also, the number of output features is changed to 2 for binary classification task.

On the other hand, the pRCC dataset has no labels while contains type I ROIs and type 2 ROIs, an initial attempt was to first obtain labels for the two type pf images respectively and train a binary classifier accordingly. More specifically, the steps are: (1) Use a pretrained Very Deep Convolutional Network (VGG16) as feature extractor to turn pRCC images into vectors with 4096 dimensions. (2) Apply principal component analysis (PCA) to reduce the feature vectors into 100-dimensional vectors. (3) Apply K-Means with K = 2 to cluster the vectors to gain class labels. (4) Train a binary classifier using the same approach as for CAM16 dataset.

Another approach attempted to incorporate the pRCC dataset as additional information was to train a convolutional autoencoder with a 512-dimensional latent space learning to reconstruct pRCC images. The autoencoder is then used as a feature extractor where the WBC images will be encoded by the encoder part and the output features are used as part of the input for the final model.

Lastly, 10% of the WBC datasets come with masks that segments out the white blood cells from their background, providing valuable guidance for the model to be more sensitive to the cell other than the background.

With models trained on the pRCC and CAM16 dataset and the preprocessed images with mask applied, we can then build the final model as shown in figure 3.

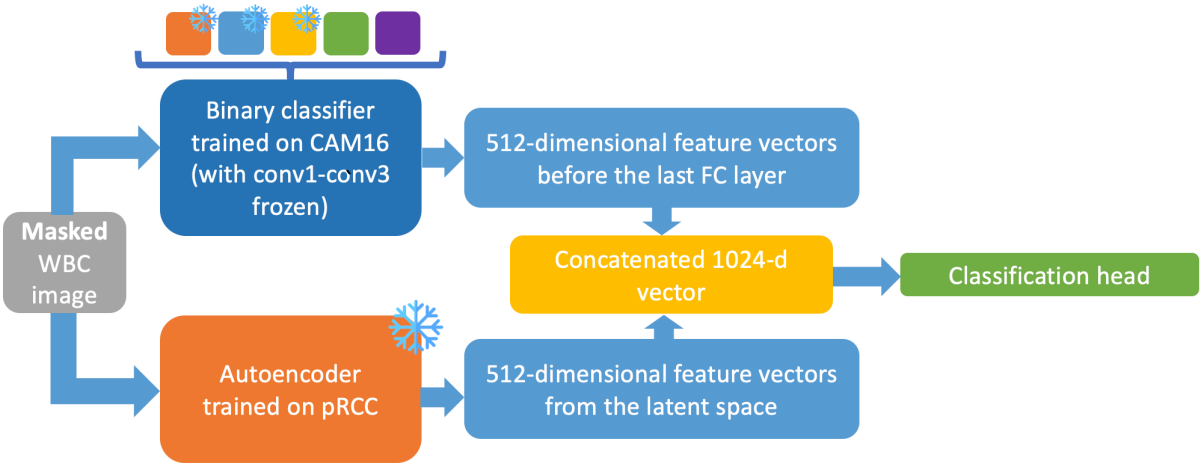

Figure 3.

Both the autoencoder (the result section briefly discusses why the binary classifier on pRCC is not chosen) and the binary classifier take an WBC image (masked if applicable) as input and output a 512-dimensional feature vector. The autoencoder is entirely frozen and is used for medium-level feature extraction only. The last two 2-layer convolutional blocks of the binary classifier are open for weight updates while the rest layers are frozen, because the binary classifier learns a high-level feature, it is not directly applicable to the WBC image, and the model still needs to learn more intricate features from the training set.

The resultant feature vectors from both models are concatenated to form a 1024-dimensional vector which will then be the input to the classification head of output dimension 5.

3. Results

The results are shown in the below tables.

Percentage of dataset used	Best test accuracy in 30 epochs (%) **without** additional information	Best test accuracy in 30 epochs (%) **with** additional information
100%	**97.90**	97.60
50%	**97.50**	97.20
10%	**95.10**	95.00
1%	80.00	**80.80**

Table 1. Training on WBC datasets

Approach used	Accuracy on classification of type I and type II (%)
VGG + K-means + ResNet	91.20

Table 2. Training on pRCC dataset

Approach used	Best test accuracy in 20 epochs (%)
ResNet	92.13

Table 3. Training on CAM16 dataset

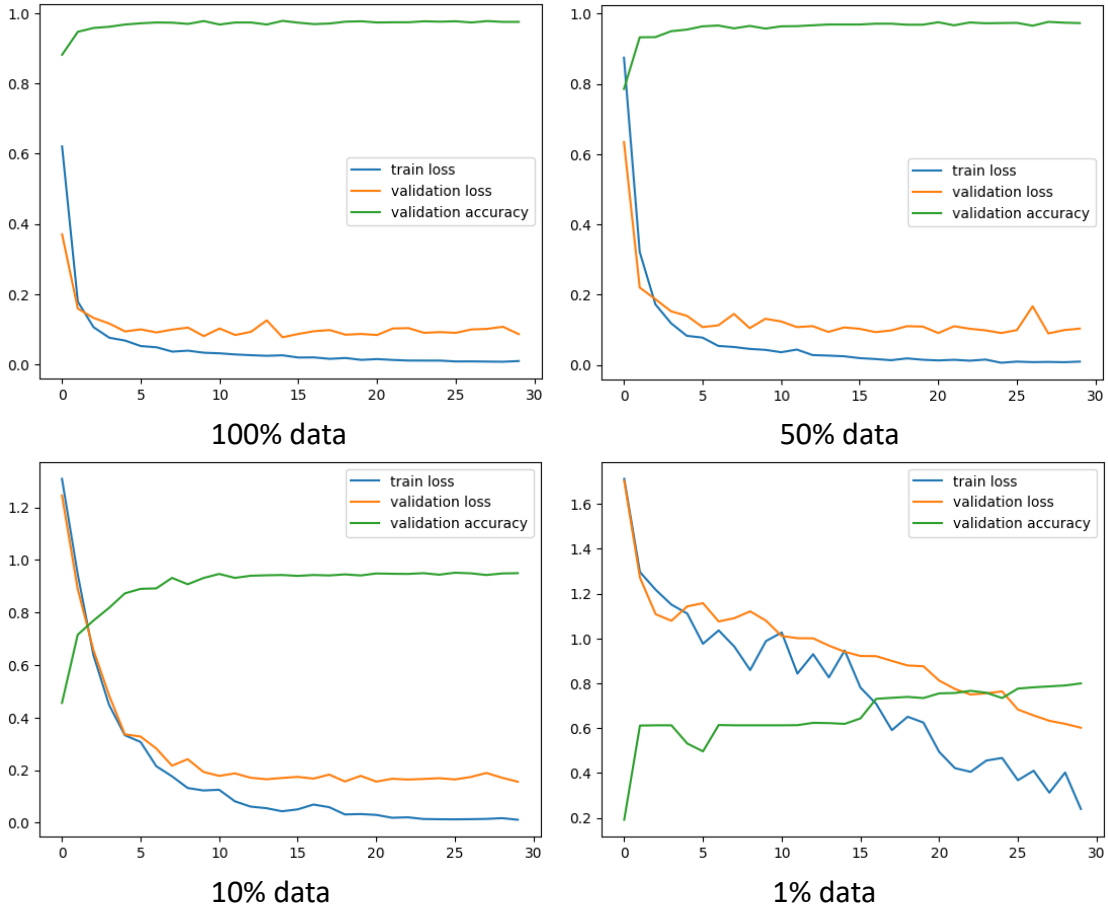

Figure 4. From left to right and top to bottom, the images show plots for training with 100%, 50%, 10% and 1% WBC unmasked data without any additional information respectively.

It can be seen from Figure 4. that as the size of dataset used decreases, the learning starts at a higher loss value and model convergence is less stable. Test accuracy stays at the same level

when using 100% and 50% data, slightly decreases when using 10% data, and significantly drop by about 15% when using only 1% training data. This is expected since in the 1% training data only a few training examples are given for classes like Basophil and Monocyte, which makes it hard to learn a good representation of certain types of WBC. While for the rest of the WBC datasets, the generalization ability present in the pretrained ResNet-18 allows the model quickly adapts to pattern of white blood cells and learn good representation even with only 10% training data.

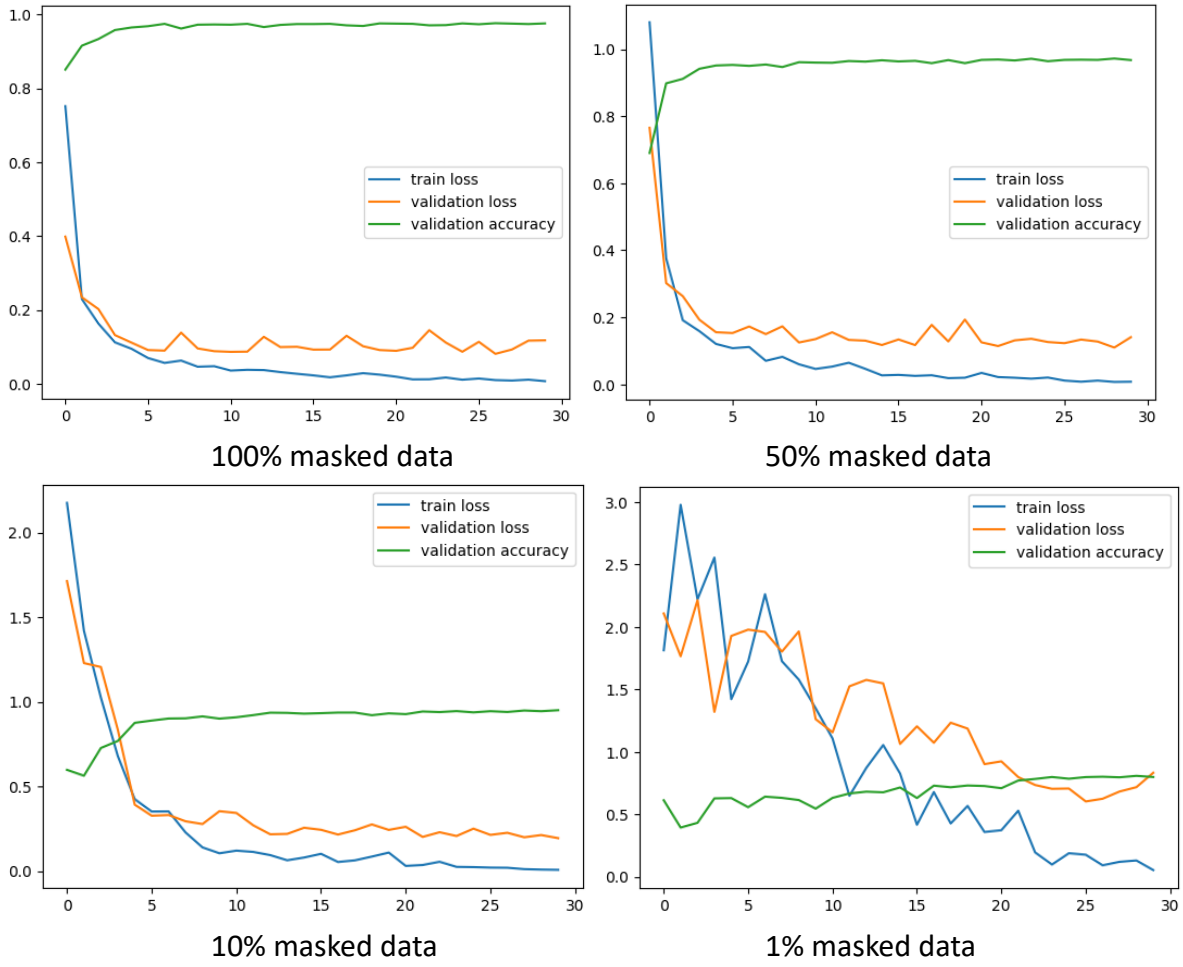

Figure 5. From left to right and top to bottom, the images show plots for training with 100%, 50%, 10% and 1% WBC masked data with additional information respectively.

From Figure 5. and Table 1., we can tell that adding additional information achieved about the same performance as compared to training on WBC datasets along. All models start with higher loss and the convergence looks less smooth probably due to the inherent difference in data distribution between WBC and the other two datasets. In particular, the training process for 1% masked data seem quite volatile, this is because insufficient data is provided for learning features unique to WBC cells and there might be some patterns common to all three medical image datasets that prevent the model from taking much weight updates to adapt to the WBC dataset.

Now we discuss why the clustering method to make label for pRCC dataset is not chosen as part of the model utilizing additional information. First, the clustering suggested two clusters of size 610 and 809 respectively, which is close to the actual number (870 type 1 ROIs and 547

type 2 ROIs mentioned in the first version of project description, the actual pRCC dataset used here has 2 more images). However, mislabelled samples still cause the binary classifier to learn the wrong knowledge. Moreover, type 1 and type 2 pRCC images have subtle differences in the cellular and cell-layer level patterns, which is hard to be captured. The autoencoder, by trying to reconstruct the image, can learn more intricate details and therefore has higher performance (when used the pRCC binary classifier, the accuracy for 1% masked data is about 73% only) when used in the model in Figure 3.

4. Conclusion

In this project, we developed two models with and without incorporating additional information that are trained on 100%, 50%, 10%, and 1% WBC datasets. The key conclusions/insights are as follows:

- Although trained on ImageNet containing natural images, pre-trained ResNet-18 demonstrated great generalization ability and adaptability on medical images.
- Pre-training with additional medical images does not seem to help with the training on WBC datasets due to 2 possible reasons. Firstly, the pre-trained network architecture is not sufficiently complex to fully capture patterns in pRCC and CAM16, when use as feature extractor for the WBC image dataset, they capture only general textures slight which is already done by the first few layers of the pre-trained ResNet, although the extracted features are adapted towards medical image, they are not intricate enough to boost the classification performance on WBC. Secondly, the three datasets are all of difference scale, our WBC datasets has the smallest scale and patterns from higher-level images related to areas like cell layer is not helpful in distinguishing different white blood cells.
- Lastly, the outcomes of this project suggest limitation of CNN-architecture in dealing with small subtle differences in large-size images (for example, pRCC) especially in the medical fields. Sometimes, spatial relationships such as relative position of cells are important in understanding medical images and therefore models with more sophisticated architecture like vision transformer could be used to tackle this task. Similarly, using autoencoder for pRCC might not be a good choice since the mean squared error loss function used in this project does not incentivize the model to learn robust features, novel objective like masked image modelling would a better choice for forcing model to focus on areas that really distinguish images from different classes.
- Future work can focus on applying novel architectures directly on classification of white blood cells and different ways of incorporating additional information.

5. Comments on the trained model files

Model files without using additional information are:
resnet_wbc_100.pt, resnet_wbc_50.pt, resnet_wbc_10.pt, resnet_wbc_1.pt

Model files using additional information are:
combinedNet_wbc_100.pt, combinedNet_wbc_50.pt,
combinedNet _wbc_10.pt, combinedNet _wbc_1.pt

Pretrained models on pRCC are:
pRCC_autoencoder.pt, pRCC_clustered_resnet.pt. CSV files used in the binary classifier approach is also given for reproducibility as prcc_feature.csv and prcc_cluster_label.csv.

Pretrained models on CAM16 are:
resnet_cam16.pt