CLASSIFYING DIFFERENT TYPES OF BLOOD CELLS

Yanran Li
Chan. School of Public Health
Harvard University
yanranli@hsph.harvard.edu

November 24, 2020

Classifying different types of blood cells and counting different blood cell are important indicators important for many diagnostic tests. Our task will use Keras and Sklearn to automatically classify 4 main types of subtype blood cells in each blood sample image and do some counting among them. Finally, we will use confusion matrix to see the models' performance and give some highlights on future work.

1 Introduction

1.1 Background

In our study, we mainly considered 4 different types of blood cells: Eosinophil, Lymphocyte, Monocyte, and Neutrophil, which are white blood cells from people's blood cells. White blood cell (WBCs) counting is an important indicator of health and is important for many diagnostic tests. Currently, doctors utilize expensive automated counters like flow cytometers, or manually count blood cells on a microscope slide. Therefore, providing an automated way to detect and count WBCs would be advantageous. Detecting the WBCs is the first step for achieving this goal.

Hematocyte / Blood Cell Blood cells are produced through hematopoiesis and found mainly in the blood. As we considered people's blood in this study: blood cells mainly contain the following three types: red blood cells, white blood cells as well as platelets. Blood cells account for about 45Among them, white blood cell (WBCs) counting is an important indicator of health and is important for many diagnostic tests. Currently, doctors utilize expensive automated counters like flow cytometers, or manually count blood cells on a microscope slide. Therefore, providing an automated way to detect and count WBCs would be advantageous. Detecting the WBCs is the first step for achieving this goal.

There are three types of WBC—lymphocytes, monocytes, and granulocytes—and three main types of granulocytes (neutrophils, eosinophils, and basophils) [1]. Their classification are shown from fig.1. With 12,500 augmented

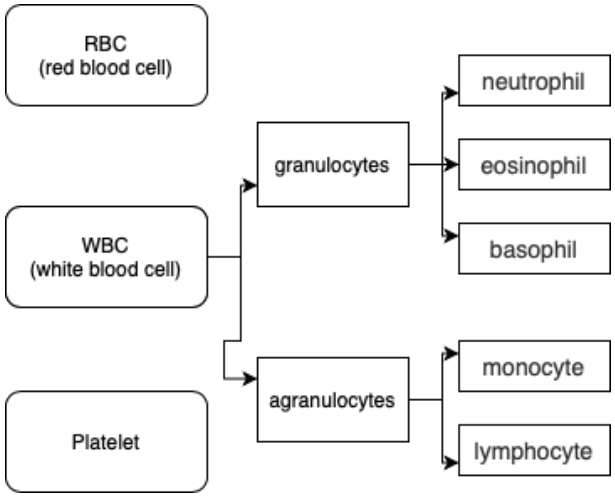

Figure 1: Blood Cells' Classification

images of blood cells paired with subtype labels (Basophil vs Eosinophil vs Lymphocyte vs Monocyte vs Neutrophil) from our dataset, we want to automatically classify each image according to the subtype of cells within it.

Mononuclear v.s. Polynuclear To help implement machine learning in the following, we see more details about the differences between subtypes of the blood cells. It makes sense that granulocytes may bring some confusing. Seeing the 3 subtypes' computer models from Fig.2-4, we've found that Eosinophi and Neutrophi both have irregular nuclei while Basophi has a more complete nucleus. To make our following machine learning easier, we labelled the subtypes (we didn't consider Basophil in this study):

- Mononuclear (Lymphocyte vs Monocyte)
- Polynuclear (Neutrophil + Eosinophil)

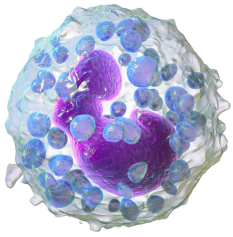

Figure 2: Basophil

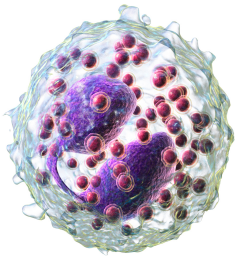

Figure 3: Eosinophil

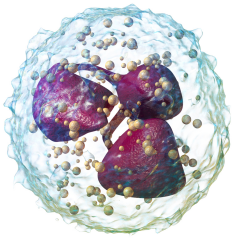

Figure 4: Neutrophil

1.2 Data

We use the "Blood Cell Images" from Kaggle `https://www.kaggle.com/paultimothymooney/blood-cells`, which contains 12,500 augmented images of blood cells (JPEG) with accompanying cell type labels (CSV). There are approximately 3,000 images for each of 4 different cell types grouped into 4 different folders (according to cell

type). The cell types are Eosinophil, Lymphocyte, Monocyte, and Neutrophil. We also use "BCCD Dataset" from `https://github.com/Shenggan/BCCD_Dataset`, which is a small-scale dataset for blood cells detection. And we contain the reference dataset from `https://bbbc.broadinstitute.org/BBBC045`.

2 Implementation

2.1 Data Preparation

With 12,500 augmented images averagely assigned to 5 different subtypes in 5 folders, we read them from their folders and gave them 2 labels stored in dictionaries:

- 1:'Neutrophil',2:'Eosinophil',3:'Monocyte',4:'Lymphocyt'
- 0:'Mononuclear',1:'Polynuclear'

Then, we used "opencv" to load pictures and transformed them into arrays with normalized 0-255 by using "scipy.misc.imresize". At this time, we got our input data, whose tensor shape is (60, 80, 3), for the following CNN model(3 for RGB label).

2.2 Convolutional Neural Network

Under the frame of keras, we set the batch size to 128 and ran 30 epochs to train our model. The model we built contains 3 Convolutional layers with their activation function ReLU. During each iteration, the flattened data finally experienced the full connection layer with a softmax function to generate its predicted value. We did 2 kinds of training: one for 4 subtypes of WBC and the other, which we assumed easier for CNN to learn, Mononuclear v.s. Polynuclear.

3 Discussion

3.1 Classification for 4 Subtypes

During our training process, each epoch cost around 170 seconds and finally the training accuracy reached 88.39% at epoch28 while the test accuracy reached 86.93% at epoch27. Model's loss and accuracies trends among 30 epochs are shown in Fig.5 and Fig.6.

It makes sense that both in Model Loss Trend and in Model Accuracy Trend, the "train" owns a smoother line plot than "test". Meanwhile, the train loss decreased stably by epochs with train accuracy increased stably.

3.2 Binary Classification for Polynuclear and Mononuclear

The binary classification model ran much faster than the above one: each epoch cost around only 43 seconds to iterate.

The training accuracy reached 98.12% at epoch28 while the test accuracy reached 94.61% at epoch19. Both are presenting better than 4-types' classification. Model's loss and accuracy trends among 30 epochs are shown in Fig.7

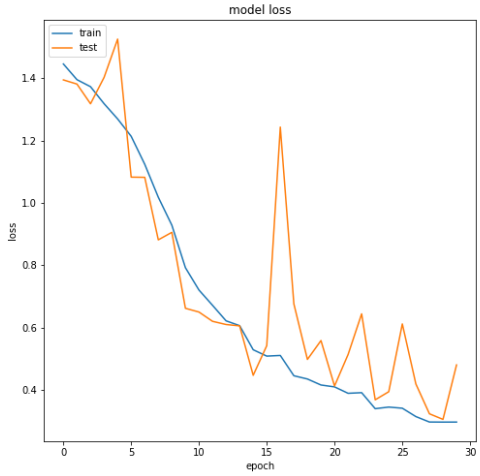

Figure 5: Model Loss Trend

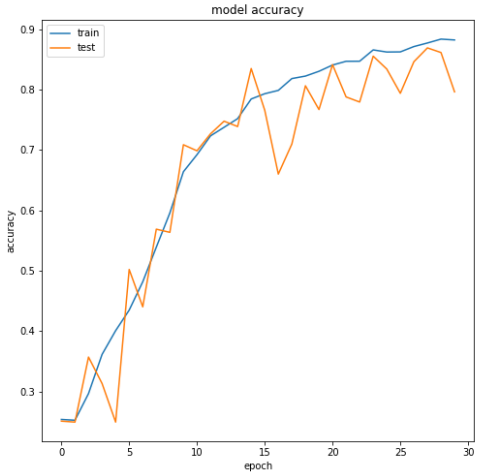

Figure 6: Model Accuracy Trend

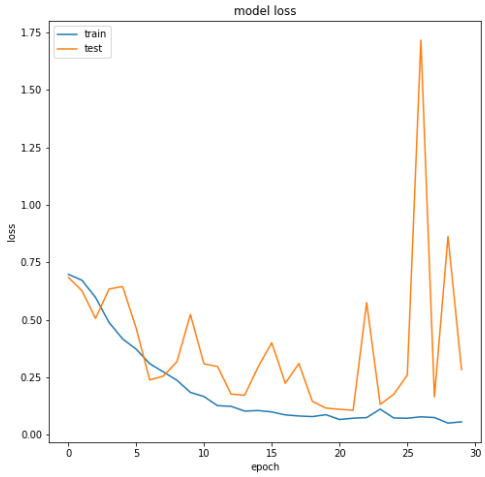

Figure 7: Model Loss Trend

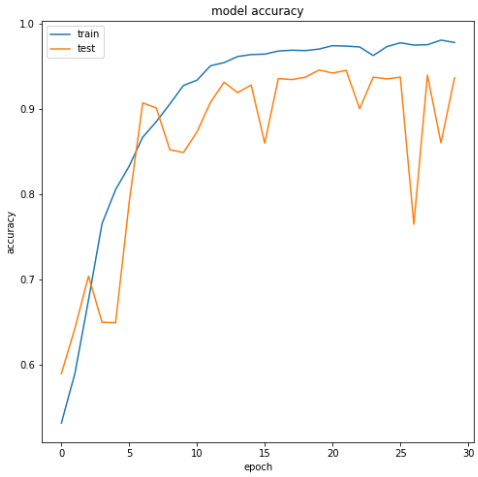

Figure 8: Model Accuracy Trend

and Fig.8. Comparing with Fig.5 and Fig.6, both the model accuracy and model loss lines of the binary classification are less stable than 4-types' classification, especially in latter epochs.

3.3 Model Summary

After checking both models' confusion matrix, we found that: with over 95% accuracy for two categories as compared to around 85% accuracy for four categories, they both work quite well than the randomly predicted probability (25%). We can state that machine learning methods like this can be improve efficiency in clinical research.

4 Future Work

Assist in Diagnosis With the classification model, we can first use centrifugal technology to clarify different types of patients' different types of blood cells and then using our model to quickly check all the cells. This can help us quickly locate some abnormal cells. It can greatly improve doctors' diagnosis efficiency and reduce the missed diagnosis and misdiagnosis.

Leukemia Leukemia is a group of blood cancers that usually begin in the bone marrow and result in high numbers of abnormal blood cells. Symptoms occur due to a lack of normal blood cells. There are four main types of leukemia—acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML)—as well as a number of less common types. Leukemias and lymphomas both belong to a broader group of tumors that affect the blood, bone marrow, and lymphoid system, known as tumors of the hematopoietic and lymphoid tissues. Its 5-year survival rate in US is only 57%, which is quite low. Certainly, early detection of abnormal blood cells and treatment can greatly increase this rate.

So far, diagnosis is typically made by blood tests or bone marrow biopsy. Sometimes a blood test can't find that the patient has leukemia, especially in the early stage or remission stage. However, we can use such models to learn confirmed patients' abnormal WBCs and help find whether the patients' blood cells had mutation occurred. Rapid localization and recognition of abnormal blood cells with such model would help the Leukemia's diagnosis and improve doctors' efficiency before the bone marrow biopsy.

References

[1] Dean L. Blood Groups and Red Cell Antigens [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005. Chapter 1, Blood and the cells it contains. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2263/