AI for Healthy vs. Acute Lymphoblastic Leukemia Cell Classification

Rambaldi Matteo, Carraro Amedeo

Abstract

Blood provides a mechanism through which nutrients, gases, and waste can be transported throughout the body and is composed of a significant number of cells suspended in a medium known as plasma. Among these, we can find erythrocytes (red blood cells), platelets and leukocytes or white blood cells.

Cells are important because they represent an easily accessible population whose morphology, biochemistry, and ecology can provide indications about a patient's general condition or clues for diagnosing diseases. It is necessary in these cases to be able to recognize normal white cells and distinguish them from pathological (diseased) cells.

The study of blood cells is primarily based on observing the presence and/or absence of their nucleus or cytoplasmic granules; other important characteristics for observation are size, nucleus size, appearance of chromatin, and cytoplasmic staining.

Lymphocytes, specific white blood cells which are the subject of this analysis, can appear both small and large. Small lymphocytes are inactive, are the same size as erythrocytes, and contain a dark nucleus that occupies almost all of the cytoplasm with a thin rim of surrounding cytoplasm, but no visible granules. Larger ones (10 - 15 microns) contain more cytoplasm, which remains basophilic, and are in an active state as B or T cells, depending on the specific signal they are able to activate.

Acute lymphoblastic leukemia (ALL) is very common in pediatric age and reaches its peak incidence between 2 and 5 years of age. It presents as a blood cancer originating from lymphocytes and is characterized by an accumulation of these cells in the blood, bone marrow, and other organs. The term "acute" defines that the disease progresses very rapidly. Specifically in ALL, an immature B or T lymphocyte undergoes a tumoral transformation: the processes that should lead to cell maturation are blocked, and it begins to replicate more rapidly, invading the blood and also reaching the lymph nodes, spleen, liver, and central nervous system.

The main risk factors for this disease are few: exposure to radiation (including for medical treatments such as radiotherapy) and to chemicals such as benzene, a natural component of petroleum, found in some pesticides and in cigarette smoke. Among the factors that cannot be controlled are pediatric age (already mentioned above) and the male gender of the subject, as well as a greater predisposition for development if the latter already has hereditary syndromes linked to genetic abnormalities.

Symptoms of this disease manifest rapidly: initially they are often nonspecific and include fatigue, loss of appetite, night sweats, and fever. Later, there is noticeable fatigue and pallor related to anemia, an increased risk of infection due to the reduction of normal white blood cells, and frequent bleeding related to platelet deficiency. Among the systemic symptoms, widespread muscle and joint pains, a sense of malaise, and weight loss are frequent. Furthermore, if it has spread to other organs, enlargement of the spleen, liver, and lymph nodes can be observed, and if the nervous system has also been reached, headaches and various neurological signs may occur.

Developing countries, such as Argentina, Chile, India, Brazil, and so on, do not have an adequate number of hospitals with modern equipment and qualified doctors. Therefore, a significant percentage of the population in these countries, especially in rural areas, is unable to avail themselves of specific and timely healthcare.

Description of the Dataset

To achieve the objective of this project, the "C-NMC Challenge" dataset is utilized, consisting of microscope images of blood smears or spots representing both healthy subjects and those affected by B-cell acute lymphoblastic leukemia (B-ALL).

The dataset comprises a total of 118 subjects, with 69 subjects diagnosed with cancer and 49 in a healthy state. The uniqueness lies in the data or images division, collected based on their function, namely the particular phase of training or validation of neural networks and the state of B-lymphoblastic leukemia cells (sick) or healthy B-lymphoid precursors (healthy cells). There are three specific sectors of use, divided as follows:

- Train Set:
  - Total Subjects: 73, ALL: 47, Healthy: 26
  - Total cells: 10,661, ALL: 7272, Normal: 3389
- Preliminary Test Set:
  - Total Subjects: 28, ALL: 13, Healthy: 15
  - Total cells: 1867, ALL: 1219, Normal: 648
- Final Test Set:
  - Total Subjects: 17, ALL: 9, Healthy: 8
  - Total cells: 2586, ALL: 1761, Normal: 825

	Phase	I	I	I	II	III
		Fold1	Fold2	Fold3		
Sub.	Sick	19	11	17	13	9
	Healthy	9	3	14	15	8
Cells	No-Healthy	2397	2418	2457	1219	-
	Normal	1130	1163	1096	648	-

References

- **ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging Select Proceedings**
  by Anubha Gupta and Ritu Gupta
- **Heterogeneity Loss to Handle Intersubject and Intra-subject Variability in Cancer**
  by Shubham Goswami, Suril Mehta, Dhruva Sahrawat, Anubha Gupta, and Ritu Gupta
- **SDCT-AuxNet: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis**
  by Shiv Gehlot, Anubha Gupta, and Ritu Gupta
- **Detection of red and white blood cells from microscopic blood images using a region proposal approach**
  by Cecilia Di Ruberto, Andrea Loddo, and Lorenzo Putzu
- **White Blood Cells Identification and Counting from Microscopic Blood Image**
  by Lorenzo Putzu and Cecilia Di Ruberto
- **Neighborhood Correction Algorithm for Classification of Normal vs . Malignant Cells**
  by Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen