AI for Healthy vs. Acute Lymphoblastic Leukemia Cell Classification

Rambaldi Matteo, Carraro Amedeo

Data Description

The dataset comprises a total of 118 subjects, with 69 subjects diagnosed with cancer and 49 in a healthy state. Inside the dataset there are no missing data, but as describe before there is an imbalanced between the number of patients/subjects diagnosed with cancer and in a healthy state. We report under the data tableaux for each phase, and looking inside that, we understood that we must apply a data-augmentation since the Train Set present the 46.60% of normal cells with respect to the 7272 cells for the ALL. As describe above the classes of this particular problem are two: B-lymphoblastic leukemia cells (sick) - ALL described by the value 1 or healthy B-lymphoid precursors (healthy cells) - NORMAL or HEM described by value 0. 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	-

All the image names follow a standard naming convention which is described below:
Cancer cell images' naming convention: **UID_P_N_C_all**

- UID_**P** $\rightarrow$ where P=1,2.... signifies the subject ID.
- UID_P_**N**: where N=1,2,3... represent the image number.
- UID_P_N_**C**: where C=1,2,3... represents the cell count.
- UID_P_N_C_all: The 'all' tag represent the class to which the cell belongs, in this case, 'ALL' or cancer class.

Similarly, the naming convention for normal (healthy) cell images is as follows: **UID_HS_N_C_hem**, where H denotes healthy/normal subject, S denotes the healthy subject's ID, N denotes the image number, C denotes the cell count, and hem tag, in the end, denotes the normal subjects' cell.

The SBILab team preprocessed these images using segmentation, image enhancement, and normalization techniques. Individual lymphocytes were segmented from blood smear images and placed in the center of them; each picture has 450×450 pixels and a black background.

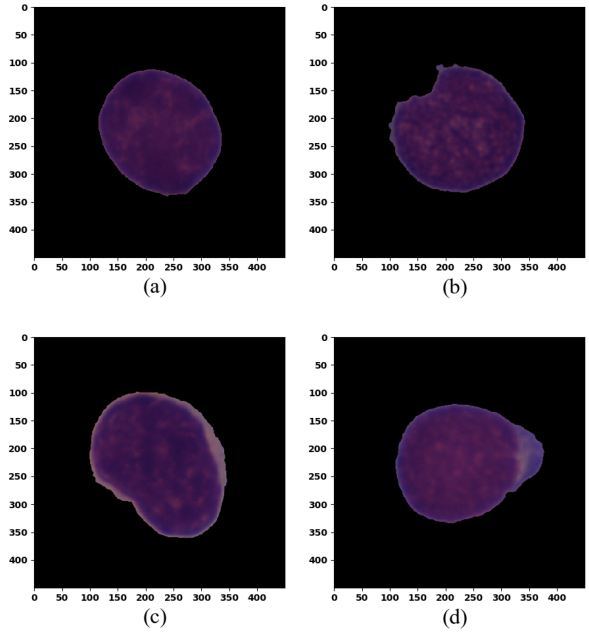

Figure 1. C-NMC 2019 dataset samples. The images (a,c) are malignant lymphocytes, and (b,d) are healthy lymphocytes.

Since that the data is already pre-processed and does not require any further processing. From those images, we applied data augmentation to balance the training and validation sets, but we don't apply that to the test images.

New images were created using and combining rotation, blurring, mirroring, shearing transformation and addition of salt-and-pepper noise.

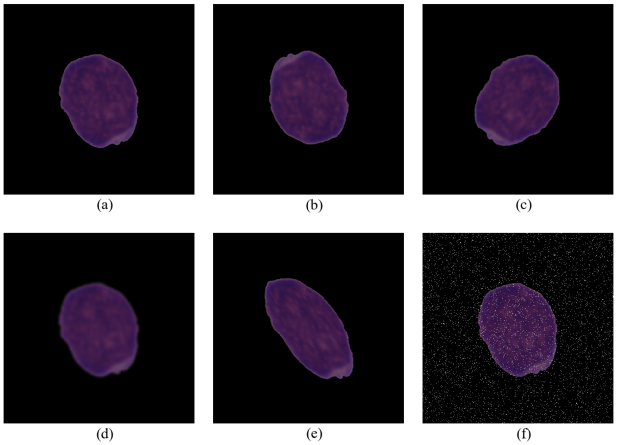

Figure 2. Examples of augmented images: (a) source image; (b) vertical and horizontal mirroring; (c) 60° clockwise rotation; (d) Gaussian blur with 17 × 17 kernel; (e) shear transformation with a factor of 0.3; and (f) salt-and-pepper noise.

For the features extraction, from each image contained in the dataset, we extracted an array of 1387 features. We used low-order statistical, textural, morphological, contour, and DCT features extracted from each lymphocyte image.

We obtained the low-order statistics from each channel of

Feature Type	Number
Low-order statistical	108
Textural	75
Morphological	20
Contour	160
DCT	1024
Total	1387

the images in both RGB and HSV formats. The textural features were calculated using the coefficients of co-occurrence matrices. These coefficients represent the different gray level combinations that occur in the image and can be used in image classification tasks. We used features obtained from the gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level dependence matrix (GLDM), gray level size zone matrix (GLSZM), and neighboring gray-tone difference matrix (GLDM). The morphological features used, indicate the general shape of a lymphocyte. We obtained the contour features from the discrete Fourier transform of the centroid distance function (CDF) of the lymphocyte. The CDF represents the distance between the lympocyte centroid and each pixel of its contour. In the end, we calculated the DCT from the lymphocyte image converted to grayscale, producing a matrix with 202.500 DCT coefficients. The size of of this matrix was the same as the number of pixels in each image (450 × 450). We mapped the coefficients to a 1D array using a zigzag scan and used only the first 1024 lowest frequency coefficients. Finally, we combined all the features into a unique vector for each sample image for the training phase and normalized all values by subtracting each value from the column's mean and dividing it by the column's standard deviation.

To achieve high-performance we are thinking about combining different lightweight classifiers into a single solution. The idea is to adopt a Neural Network as Residual Neural Network or VGG16, Support Vector Machine - SVM and Naive Bayes Classifier - NB.

References

- **GCTI-SN: Geometry-Inspired Chemical and Tissue Invariant Stain Normalization of Microscopic Medical Images**

  Anubha Gupta, Rahul Duggal, Ritu Gupta, Lalit Kumar, Nisarg Thakkar, and Devprakash Satpathy

- **Stain Color Normalization and Segmentation of Plasma Cells in Microscopic Images as a Prelude to Development of Computer Assisted Automated Disease Diagnostic Tool in Multiple Myeloma**

  Ritu Gupta, Pramit Mallick, Rahul Duggal, Anubha Gupta, and Ojaswa Sharma

- **Neighborhood Correction Algorithm for Classification of Normal vs . Malignant Cells**

  by Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen

- **Overlapping Cell Nuclei Segmentation in Microscopic Images UsingDeep Belief Networks**

  Rahul Duggal, Anubha Gupta, Ritu Gupta, Manya Wadhwa, and Chirag Ahuja

- **Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks**

  Rahul Duggal, Anubha Gupta, and Ritu Gupta

- **SD-Layer: Stain Deconvolutional Layer for CNNs in Medical Microscopic Imaging**

  Rahul Duggal, Anubha Gupta, Ritu Gupta, and Pramit Mallick