C-NMC Dataset

Aim: Classification of leukemic B-lymphoblast cells (cancer cells) from normal B-lymphoid precursors (normal cells) from blood smear microscopic images.

A dataset of cells with labels (normal versus cancer) is provided to train machine learning-based classifier to identify normal cells from leukemic blasts (malignant/cancer cells). These cells have been segmented from the microscopic images. These images are representative of images in the real-world because these contain some staining noise and illumination errors, although these errors have largely been fixed by us via our own in-house method of stain color normalization.

The ground truth has been marked by an expert oncologist.

This dataset was also used for our IEEE ISBI 2019 conference challenge: *Classification of Normal vs Malignant Cells in B-ALL White Blood Cancer Microscopic Images.* The challenge is available here:

https://biomedicalimaging.org/2019/challenges/

https://competitions.codalab.org/competitions/20429

Description of dataset

The folder contains data arranged in three folds. For example, if Fold1 contains full data from subject IDs 1,2,3,4,5 then Fold2 contains full data from subject IDs 6, 7, 8, 9,10. No two splits overlap in terms of subject data i.e. subject ID found in Fold1 will only be present in Fold1.

- Fold1:
    - all
        - Image1, Image2, ...
    - hem
        - Image3, Image4, ...
- Fold2:
    - all
        - Image5, Image6, ...
    - hem
        - Image7, Image8, ...
- Fold3:
    - all

        - Image9, Image10, ...
    - hem
        - Image11, Image12, …

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** -> 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. (More than one cell can be found in a particular microscopic image)
- 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 dataset contains a total of 118 individual subjects, distributed as follows:

- ALL (cancer) subjects: 69
- Normal subjects: 49
- **Train set composition:**
    - Total subjects: 73, ALL: 47, Normal: 26
    - Total **cells**: 10,661, ALL: 7272, Normal: 3389
- **Preliminary test set composition:** Total subjects: 28, ALL: 13, Normal: 15
    - Total Cells: 1867, ALL: 1219, HEM: 648

- **Final test set composition:** Total subjects: 17, ALL: 9, Normal: 8

    - Total Cells: 2586

Please note that the ground truth labels of the final test set are not provided. The results of classification should be tested on this dataset and checked at the leaderboard of the codalab challenge to know the comparative performance with the world teams. The evaluation metric is weighted f1 score. The process to check results is as below-

1. Please register at the codalab challenge page by clicking the button of “Sign in” at the below page-

   https://competitions.codalab.org/competitions/20429#participate

2. Now, you can submit your results of the test_final_phase data for checking. You will be able to see your comparative performance.

Who would like to work on this problem?

This problem is very challenging because as stated above, morphologically, the two cell types appear very similar. The ground truth has been marked by the expert based on domain knowledge. Also, with our efforts in the past two years, we have also recognized that the subject level variability also plays a key role and as a consequence, it is challenging to build a classifier that can yield good results on prospective data. Anyone deeply interested in working on a challenging problem of medical image classification via building newer deep learning/machine learning architectures would, in our opinion, come forward to work on this challenge.

What general pre-processing steps will be performed?

The data is already preprocessed and does not require any further processing. However, participants are free to apply any further processing techniques, if required.

Please cite the following papers if this dataset is used for any publication:

1. Anubha Gupta, Rahul Duggal, Ritu Gupta, Lalit Kumar, Nisarg Thakkar, and Devprakash Satpathy, “GCTI-SN: Geometry-Inspired Chemical and Tissue Invariant Stain Normalization of Microscopic Medical Images,”, under review.
2. Ritu Gupta, Pramit Mallick, Rahul Duggal, Anubha Gupta, and Ojaswa Sharma, "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," 16th International Myeloma Workshop (IMW), India, March 2017.
3. Rahul Duggal, Anubha Gupta, Ritu Gupta, Manya Wadhwa, and Chirag Ahuja, “Overlapping Cell Nuclei Segmentation in Microscopic Images UsingDeep Belief Networks,” Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), India, December 2016.
4. Rahul Duggal, Anubha Gupta, and Ritu Gupta, “Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks,” CME Series on Hemato-

Oncopathology, All India Institute of Medical Sciences (AIIMS), New Delhi, India, July 2016.

5. Rahul Duggal, Anubha Gupta, Ritu Gupta, and Pramit Mallick, "SD-Layer: Stain Deconvolutional Layer for CNNs in Medical Microscopic Imaging," In: Descoteaux M., Maier-Hein L., Franz A., Jannin P., Collins D., Duchesne S. (eds) Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017, MICCAI 2017. Lecture Notes in Computer Science, Part III, LNCS 10435, pp. 435–443. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-66179-7_50.