Multilabel Leukemia classification machine

Section 3 DATASCI 281 Computer Vision
Dylan Jin, Fabian Riquelme, Luc Robitaille, Dmitri Zadvornov

Introduction and Problem Statement

It is estimated that nearly 60,000 Leukemia blood cancers are diagnosed annually in the United States (Leukemia and Lymphoma Society, 2021). A common diagnosis tactic for Leukemia is a peripheral blood smear - this test is minimally invasive, only requiring a blood draw. Preparing and classifying a Peripheral Blood Smear is a highly skilled and laborious task that requires the attention of trained lab technicians and medical experts (American Cancer Society, 2018). Generally speaking, instances of cancer can be classified into stages based on their progression. Cancers caught early in their progression are far more likely to be treatable (He et al., 2022). Accurate diagnosis automation creates the opportunity for more regular testing, as the cost of tests is driven down by automation. Furthermore, as costs decline, testing frequency can increase, allowing for a higher sampling rate of individuals. Higher sampling rates will increase the likelihood of early detection and survival rate of the Leukemia family of cancers. We propose a project to perform multilabel classification on Acute Lymphoblastic Leukemia (ALL) peripheral blood smear data to classify ALL and the stage present in the sample. Furthermore, we plan to study the effectiveness of manually created feature filters with traditional Machine Learning models (e.g. Support Vector Machines) versus more advanced techniques exhibited in Deep Learning to examine the performance differences when labeled data is not abundant.

Research proposal

We hypothesize that, given a limited dataset size, standard feature-extraction-based approaches would produce superior outcomes since models based on CNN extraction require a considerable amount of data to be fully trained. For our CNN-based model, we will employ EfficientNet-B3 as our benchmark. In addition, our feature extraction-based model will involve three traditional methods - LBP (Singhal & Singh, 2014), Sobel Edge Detection, and HoG - and one complex-based approach (Alex-Net). Features will be implemented in traditional Machine Learning classification models: SVM and Decision Tree. In both methods, we will obtain multilabel classifications, with the primary metrics being accuracy and F1 scores.

Literature review

Analyzing recent research work, we found that model performance is driven, among other things, by the feature extraction method, feature filtering (reduction) method, and selection of the classification model (He et al., 2022).

One of the recent State of the Art (SOTA) methods utilized CNN models based on a novel, more efficient approach called EfficientNet-B0-B7 that have greater classification accuracy while requiring fewer parameters. EfficientNet-B3 enabled dynamic classification of ALL cells and resulted in an accuracy score of 97.57% and F1-score of 98.22% (Abd El-Ghany, 2023). Other alternative methods not based on CNN-based classification utilize SVM with high accuracy

scores. However, they do not perform as well as the CNN-based models, with an accuracy of around 96% and an F1 score of 96.22% (Sahlol et al., 2020).

Typical feature extraction methods focus on traits such as color, shape, and texture (Hegde et al., 2020). Some works also include statistical features such as skewness, variance, or gradient matrix (Patel et al., 2015). More recently, CNN-based extraction methods are gaining popularity, such as AlexNet, CaffeNet, and VGG-f, among others (Vogado et al., 2018). Feature filtering methods include PCA (Das et al., 2020), LDA, and SESSA (Sahlol et al., 2020).

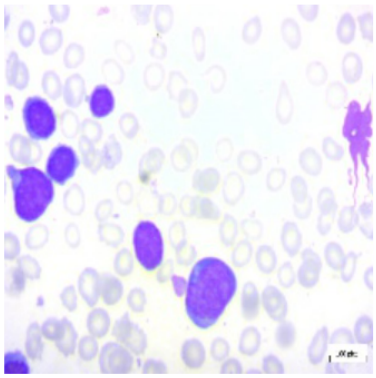
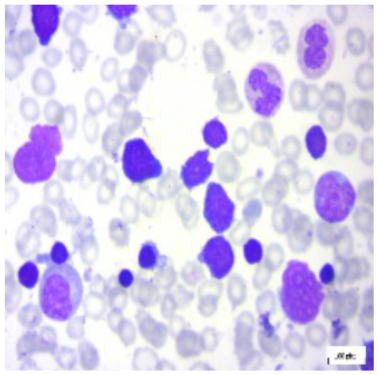
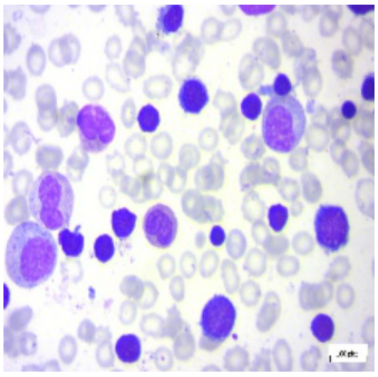
Classification includes traditional methods like DT, SVM, KNN, and DL transfer learning methods based on the models such as VGG, EfficientNet, InceptionNet, and others.

Data

For this project, we will use a dataset containing 3562 peripheral blood smear images from 89 patients suspected of ALL stained by laboratory staff with cytochemical dyes. Each image is split into ALL subtypes and benign cells. The dataset comprises 504 images of healthy cells with benign diagnosis, 985 images of Early Pre-B cells, 963 images of Pre-B cells, and 804 images of Pro-B malignant lymphoblasts. Classification of cell types and subtypes was performed by a specialist using flow cytometry instrumentation.

Each Raw image is RGB formatted with 224 x 224 pixels. The images were taken with the same Zeiss camera on a microscope with 100x magnification and saved as a JPG file. Each image contains one or more stained lymphocyte cells. The cytoplasm of lymphocytes reacts with a light blue color and the nucleus with a deep blue-violet color.

Link to dataset: https://www.kaggle.com/datasets/mehradaria/leukemia

Operational Pipeline

1. Image Processing
    1.1. Image Smoothing
        1.1.1. Adaptive Histogram Equalization - Image Sharpening
    1.2. Color Normalization
        1.2.1. Histogram Normalization - Image Sharpening
    1.3. Data Augmentation
        1.3.1. Increase sample size through image transformations
        1.3.2. Dataset balancing

2. Image Segmentation
    2.1. Threshold Segmentation - Separation of cells from background
    2.2. Clustering Segmentation - Unsupervised separation of cells
    2.3. Watershed Segmentation - Morphology segmentation based on contours

3. Feature Extraction
    3.1. Texture Features
        3.1.1. Histogram of Gradients (HoG) - Texture / Contour Features
    3.2. Shape Features
        3.2.1. Sobel Edge Detection
    3.3. Color Features
        3.3.1. Local binary patterns (LBP)

4. Classification
    4.1. Traditional Classification
        4.1.1. Support Vector Machines - Standard supervised learning algorithm for image classification
        4.1.2. Decision Tree
    4.2. Deep Learning
        4.2.1. EfficientNet family of architectures

References

1. Leukemia and Lymphoma Society. (2021). Facts and Statistics. https://www.lls.org/facts-and-statistics/facts-and-statistics-overview
2. American Cancer Society. (2018, October 17). Tests for Acute Lymphocytic Leukemia (ALL). https://www.cancer.org/cancer/acute-lymphocytic-leukemia/detection-diagnosis-staging/how-diagnosed.html
3. Wenbin He, Ting Liu, Yongjie Han, Wuyi Ming, Jinguang Du, Yinxia Liu, Yuan Yang, Leijie Wang, Zhiwen Jiang, Yongqiang Wang, Jie Yuan, & Chen Cao. A review: The detection of cancer cells in histopathology based on machine vision. Computers in Biology and Medicine, Volume 146, 2022, 105636. https://doi.org/10.1016/j.compbiomed.2022.105636.
4. Luis H.S. Vogado, Rodrigo M.S. Veras, Flavio. H.D. Araujo, Romuere R.V. Silva, Kelson R.T. Aires. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. Engineering Applications of Artificial Intelligence, Volume 72, 2018, Pages 415-422, ISSN 0952-1976. https://doi.org/10.1016/j.engappai.2018.04.024.
5. Sahlol, A.T., Kollmannsberger, P. & Ewees, A.A. Efficient Classification of White Blood Cell Leukemia with Improved Swarm Optimization of Deep Features. Sci Rep 10, 2536 (2020). https://doi.org/10.1038/s41598-020-59215-9
6. Abd El-Ghany S, Elmogy M, El-Aziz A. Computer-Aided Diagnosis System for Blood Diseases Using EfficientNet-B3 Based on a Dynamic Learning Algorithm. Diagnostics (Basel). 2023 Jan 22;13(3):404. doi: 10.3390/diagnostics13030404. PMID: 36766509; PMCID: PMC9913935.
7. Hegde RB, Prasad K, Hebbar H, Singh BMK, Sandhya I. Automated Decision Support System for Detection of Leukemia from Peripheral Blood Smear Images. J Digit Imaging. 2020 Apr;33(2):361-374. doi: 10.1007/s10278-019-00288-y. PMID: 31728805; PMCID: PMC7165227. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7165227/
8. Mehrad Aria, Mustafa Ghaderzadeh, Davood Bashash, Hassan Abolghasemi, Farkhondeh Asadi, and Azamossadat Hosseini, “Acute Lymphoblastic Leukemia (ALL) image dataset.” Kaggle, (2021). DOI: 10.34740/KAGGLE/DSV/2175623.
9. Das, P.K., Jadoun, P., & Meher, S. (2020). Detection and Classification of Acute Lymphocytic Leukemia. 2020 IEEE-HYDCON, 1-5.
10. Nimesh Patel, Ashutosh Mishra. Automated Leukaemia Detection Using Microscopic Images, Procedia Computer Science, Volume 58, 2015, Pages 635-642, ISSN 1877-0509 https://doi.org/10.1016/j.procs.2015.08.082.
11. V. Singhal and P. Singh, "Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia," 2014 Twentieth National Conference on Communications (NCC), Kanpur, India, 2014, pp. 1-5, doi: 10.1109/NCC.2014.6811261.