Lymphocytosis Classification Challenge

Joachim COLLIN JOACHIM.COLLIN@ELEVES.ENPC.FR

Bastien LE CHENADEC BASTIEN.LE-CHENADEC@ELEVES.ENPC.FR

Abstract

1. Introduction

Lymphocytosis is a common hematologic abnormality characterized by an increase in the absolute concentration of lymphocytes to more than 4000 lymphocytes/microL for adult patients (Hamad and Mangla, 2023). This condition can arise from various sources, including reactions to infections, drugs, or stress, or it may indicate a lymphoproliferative disorder, which is a type of cancer involving abnormal proliferation of lymphocytes. Clinicians typically diagnose lymphocytosis by assessing personal data such as medical history, symptoms, medication lists, and through a blood test to measure lymphocyte levels. However, additional tests may be necessary to confirm the cause of lymphocytosis and determine an appropriate treatment plan. Each condition associated with lymphocytosis presents its own set of symptoms and treatment options. While the diagnosis process is efficient, it suffers from poor reproducibility, and the additional tests required can be costly and time-consuming (Sahasrabudhe et al., 2021). Being able to distinguish more accurately reactive from malignant lymphocytosis patients is challenging and would lead to a better identification of patients requiring additional testing.

2. Architecture and methodological components

2.1. Data preprocessing

As a first step in this challenge, we analyze the data distributions to make correct preprocessing and architectural choices. In figure 1, we observe that the training and test datasets have similar enough distributions of their features (taking into account the small number of samples 163 training samples and 42 testing samples).

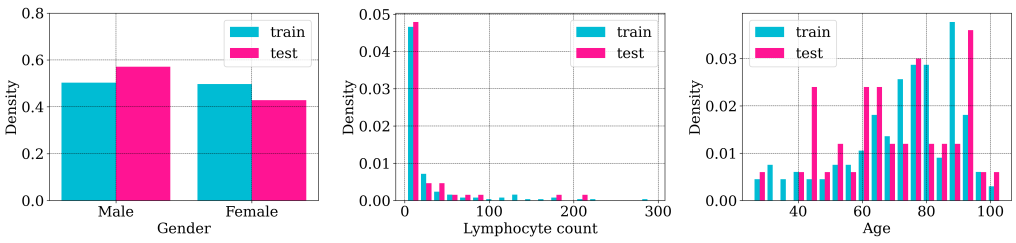

Figure 1: Histograms of the training and test datasets.

On figure 2, we observe the distribution of the positive and negative classes in the training dataset. While gender does not seem to be a good predictor of the negative (reactive) or positive (malignant) class, a high lymphocyte count seems to be a good indicator of malignant lymphocytosis. The malignant nature is also more prevalent in older patients.

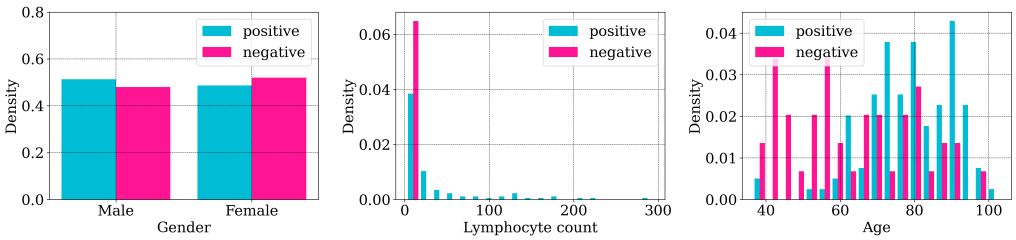

Figure 2: Histograms of the positive and negative classes in the training dataset.

The imbalance between the positive and negative classes ([TODO HOW MUCH]) justifies the use of a stratified split of the training dataset between training and validation sets, i.e. we split the dataset in a way that the proportion of positive and negative classes is the same in the training and validation sets. We also normalize the lymphocyte count and age features to be between 0 and 1.

3. Model tuning and comparison

4. Conclusions

References

Hussein Hamad and Ankit Mangla. Lymphocytosis, 7 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK549819/.

Mihir Sahasrabudhe, Pierre Sujobert, Evangelia I. Zacharaki, Eugenie Maurin, Béatrice Grange, Laurent Jallades, Nikos Paragios, and Maria Vakalopoulou. Deep Multi-Instance learning using Multi-Modal data for diagnosis of lymphocytosis. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2125–2136, 6 2021. doi: 10.1109/jbhi.2020.3038889. URL https://doi.org/10.1109/jbhi.2020.3038889.