Challenge: Lymphocytosis classification

Matteo MARENGO[1] MATTEO.MARENGO@ENS-PARIS-SACLAY.FR
Manal MEFTAH[1] MANAL.MEFTAH@ENS-PARIS-SACLAY.FR
[1] *ENS Paris-Saclay, Gif-sur-Yvette, MSc MVA, 2023-2024*

Abstract

This two-month challenge highlighted the complexities of working with medical data, particularly due to issues such as imbalanced datasets and limited data availability. Leveraging both imaging and clinical data, we developed and assessed a **Multi-Instance Learning pipeline** designed to classify patients based on their type of lymphocytosis. Our efforts focused on optimizing feature extraction, ensuring balanced training, and effectively aggregating embeddings to improve classification accuracy.

Keywords: Multiple Instance Learning, Segmentation, Autoencoders, MLP, Transfer Learning, Unbalanced data

1. Introduction

1.1. Problem Statement

Lymphocytosis, a condition characterized by an increased number of lymphocytes in the blood, can be indicative of various conditions, ranging from benign reactions **(termed reactive and labeled 0)** to serious lymphoproliferative disorders such as cancer **(termed tumoral and labeled 1)**. The current clinical practice involves **visual microscopic examination** of blood cells and integration of **clinical attributes** for diagnosis. This approach is quick and cheap, but not always consistent and sometimes needs further testing for a definite diagnosis. The goal of this paper is to create a reliable, automated method to help doctors decide who needs more tests and improve how patients are diagnosed.

1.2. Data

Concerning our data, we have at our disposal **163 patients in the train set and 42 patients in the test set**. Regarding the train set, there are **50 patients labeled as reactive and 113 patients labeled as tumoral**. Additionally, it is worth noticing that for each patient, we do not have the same number of images. Therefore, this represents in total for the train set **10861 images for the tumoral patients and 2592 reactive patients**. It is a ratio of 0.24. These unbalanced classes are important to notice as they might impact our training, more information on it can be found in the Appendix A. In addition to these images, we have clinical information on the patients. More precisely, there is **the date of birth, the gender, and the lymphocyte count (in nb/L)**. We worked using Kaggle GPU P100. All the code can be found in the dedicated file. Efforts have been made to make it as reproducible as possible. The final score on Kaggle will be assessed using Balanced Accuracy.

© CC-BY 4.0, M. MARENGO & M. MEFTAH.

2. Architecture and methodological components

2.1. Models based only on the clinical data

2.1.1. Data exploration

First steps of the challenge was to do a data exploration of the clinical data at our disposal. Rapidly, we noticed that there were **real differences between reactive and tumoral patients**. Indeed, the mean lymphocyte count for reactive patients was $5.01 \times 10^9/L$ where it was $35.90 \times 10^9/L$ for tumoral ones. Same for the year of birth, the mean one was 1965.7 for reactive patients vs 1944.8 for tumoral ones.

Data	MIN	MAX	MEAN	STD
LYMPH COUNT REACTIVE	4.01	7.68	5.01	1.00
YOB REACTIVE	1927	1998	1965.75	19.77
LYMPH COUNT TUMORAL	2.28	295.0	35.90	53.57
YOB TUMORAL	1921	1987	1944.82	12.04

Table 1: Statistics on the Lymphocyte count and on the year of birth for reactive vs tumoral patients

2.1.2. Threshold method

One initial approach we considered was the application of **basic thresholds**—specifically, one on the year of birth and another on the lymphocyte count. This method was anticipated to yield effective results, given that the distributions of reactive and tumoral patients appear quite distinct when considering these parameters. Moreover, these particular data points **have been recorded by medical professionals**, adding a level of reliability to them. We have conducted experiments to adjust the threshold values to observe the changes in the predictions for reactive versus tumoral patients.

2.1.3. Machine Learning approaches

To go further this threshold strategy, we also applied standard classifiers, such as Support Vector Machines and Decision Trees, to the clinical dataset following the methodology recommended by (Sahasrabudhe et al., 2020). Our exploration then extended to ensemble learning techniques, notably AdaBoost and XGBoost. To assess the efficacy of these classifiers, **we employed a stratified validation set, ensuring equitable subgroup representation in both training and validation phases** for accurate model evaluation. Moreover, we optimized the classifiers' hyperparameters through **grid search** to fine-tune our models. In addition to these methods, we utilized a multi-layer perceptron with two dense layers, including adjustments in layer quantity and increased hidden size. This MLP model was further refined using **stratified k-fold cross-validation.**

2.2. Multiple-Instance Learning

2.2.1. ResNet-50 based MIL

Building upon our initial findings, we shifted our focus to a **Multiple-Instance Learning (MIL)** approach. This decision was made by the assumption that blood smear images held critical diagnostic information that could further enhance our model's accuracy. As a central approach, we used a pre-trained ResNet50 model (He et al., 2015) for feature extraction from images. Recognizing each patient's dataset as a "bag" of images, the output of the ResNet50 is a set of features for each instance within the bag. To integrate the clinical data, we encoded and normalized patient-specific information, such as age and gender, for the concatenation with the image-derived features.

Our first MIL model architecture incorporates:

- An **averaging operation** across instance features within a bag to generate a singular representative feature vector that characterizes each patient dataset.
- A **linear layer that transforms clinical data into a 128-dimensional vector**, which is concatenated with the aggregated image features.
- **A sequence of fully connected layers**, including a hidden layer with 512 units, ReLU activation, and dropout for regularization. This network has a binary output layer that predicts the patient's condition as either tumoral (1) or reactive (0).

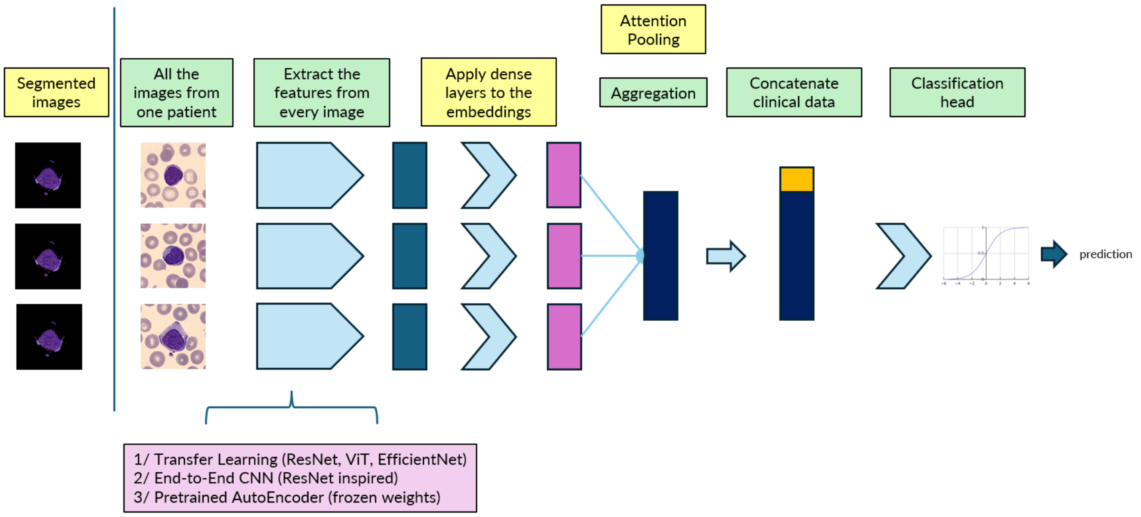
Attention Pooling
Segmented images
All the images from one patient
Extract the features from every image
Apply dense layers to the embeddings
Aggregation
Concatenate clinical data
Classification head
prediction
1/ Transfer Learning (ResNet, ViT, EfficientNet)
2/ End-to-End CNN (ResNet inspired)
3/ Pretrained AutoEncoder (frozen weights)

Figure 1: Model architecture with the different layers. Additional studies / ablation are displayed in yellow

2.2.2. Training and Validation with bag of Instances

Our dataset was split into training and validation stratified sets in **an 80-20 ratio**, useful for model generalization. To ensure personalized processing of each patient's images and clinical data, we trained the model with a batch size of **1**, considering each patient's data as an individual bag of instances. We used a **binary cross-entropy** loss function and the Adam optimization algorithm with a learning rate of 0.001 over 20 epochs. This basic MIL mode scored 0.75 on the data challenge. Some tests were made using **weighted loss** and attributing to each class the proportion of classes in the trainset to overcome unbalanced classes.

3. Model tuning and comparison

3.1. Comparison of feature extraction methods

We used different pre-trained models as feature extractors :

- **ResNet** for its ability to efficiently learn from complex images through deep residual learning.
- **EfficientNet** (Tan and Le, 2020) for its scalable architecture, optimizing for both accuracy and computational resources.
- **VGG 19** (Simonyan and Zisserman, 2015) for its depth and simplicity, providing a comprehensive baseline for feature extraction.

Model	**ResNet**	**VGG 19**	**EfficientNet**
Balanced Accuracy	0.75	0.74	0.79

Table 2: Comparison of balanced accuracies for different pretrained models submission.

We explored pretrained models which are trained on general datasets like ImageNet, posing challenges for clinical imagery analysis. To bridge this gap, we employed a **convolutional feature encoder** within a MIL framework, training it alongside the classification head for clinical relevance. Additionally, we utilized an **AutoEncoder pre-trained on image reconstruction tasks**, integrating its encoder as a static feature extractor in our model.

Feature Classification via MISVM After extracting features using ResNet50 and combining them with clinical data, we used a Multi-Instance Support Vector Machine (MISVM) (Andrews et al., 2002) to classify the labels. The MISVM algorithm differentiates between reactive and tumoral lymphocytosis by identifying the **optimal hyperplane that separates the feature space at the bag level**. The model demonstrated a balanced accuracy of 0.78 on the challenge board.

Model	**Linear Layer + FC Layers**	**MISVM**
Balanced Accuracy	0.75	0.78

Table 3: Comparison of balanced accuracy scores for models using ResNet50 with different classification approaches.

Improving the ResNet50 MIL In refining our ResNet50 MIL model, we focused on enhancing reproducibility and model performance. By setting a consistent seed, increasing epochs to 30 and saving the best model over balanced accuracy, raising the learning rate to 0.01, and adjusting the train-validation split to 75:25, **a 0.82 score was obtained**.

3.2. Results

In the Table 4, the results of our different models are shown. In addition to these models, we performed **ensemble learning** that did not succeed in improving the public score.

Model Variant	Balanced Accuracy
ResNet50 MIL	0.75
Improved ResNet50 MIL	0.82
VGG 19 MIL	0.74
EfficientNet MIL	0.79
ViT MIL	0.77
ResNet 50 MILSVM	0.78
Custom ResNet MIL Balanced Loss	0.78
Pretrained AutoEncoder MIL Balanced Loss	0.74
Threshold	0.83
XGBoost	0.79
MLP	0.79
Ensemble Learning	0.83

Table 4: Comparison of balanced accuracies between the basic and improved ResNet50 MIL models.

3.3. Additional models that did not perform well

Some tests were made using **segmented images** of the lymphocytes, and then learned the features from these segmented lymphocytes instead of the whole images. However, it does not capture the model embeddings better, as results did not increase. The first approach in MIL was to do **instance learning**, meaning that we associate each image with the class of the patient. Even if, some good submissions were obtained (0.78) it was not clinically right as a tumoral patient can have lymphocytes that are due to reactive condition for example. Another test concerned the **attention pooling** to replace mean pooling when averaging image embeddings.

4. Discussion & Conclusion

This challenge allowed us to discover the multiple instance learning field with nice architectures to implement. The work developed allowed us to do an extensive study of different models, always by respecting unbiased internal validation procedures. Avenues of improvement are the aggregation of embeddings with an attention mechanism, self-supervised learning to learn more features, data augmentation of the images, and Prompt Tuning (Zhang et al., 2023).

Acknowledgments

Thank you for Pr. Olivier COLLIOT, Pr. Maria VAKALOPOULOU and the TAs for this nice and structured course on such an interesting topic.

References

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems*, 15: 561–568, 01 2002.

Menon M. Dodd KC. Sex bias in lymphocytes: Implications for autoimmune diseases. *Front Immunol.*, 2022. doi: 10.3389/fimmu.2022.945762.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Mihir Sahasrabudhe, Pierre Sujobert, Eugénie Maurin, Beatrice 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*, PP:1–1, 11 2020. doi: 10.1109/JBHI.2020.3038889.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

Jingwei Zhang, Saarthak Kapse, Ke Ma, Prateek Prasanna, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning, 2023.

Appendix A. Distribution between reactive and tumoral patients

First, we looked at the repartition of reactive/tumoral patients and men/women in the training set as shown in Fig 2. As stated before, there are unbalanced classes between tumoral and reactive patients, however, there is an equity between men and women. Meaning that at least it will not be biased from that point of view (indeed, some studies seem to indicate that women might suffer more than men from autoimmune diseases (Dodd KC, 2022)). What is also reassuring is that there is the same number of men/women in each of the classes (tumoral/reactive) as shown in Fig 3.

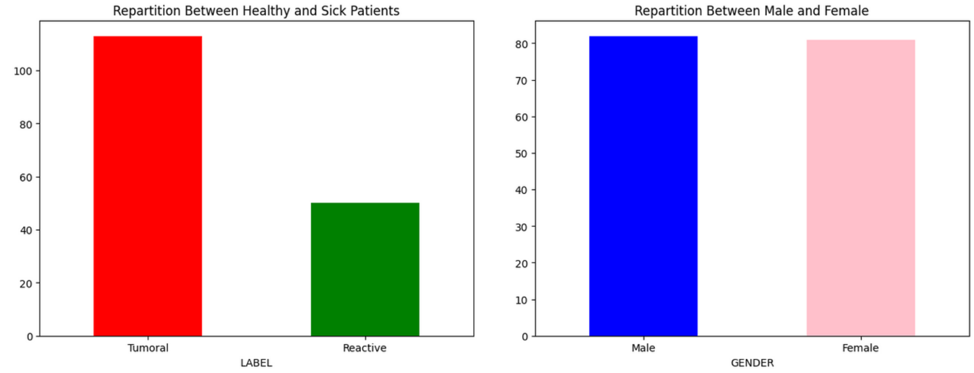

Figure 2: Distribution of the repartition of reactive/tumoral patients and men/women in the training set

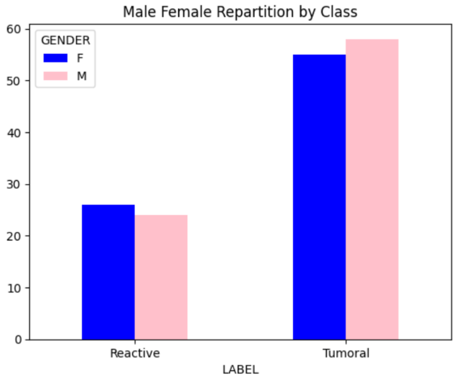

Figure 3: Distribution of the repartition of women/men in each classes

An other topic that was presented in the introduction is in addition to the unbalanced classes between reactive and tumoral patients, there is an unbalanced ratio in terms of the number of images. Indeed, there is not the same number of images for each patient. However, as seen in Fig 4 there seems to be the same distribution in the train/test set.

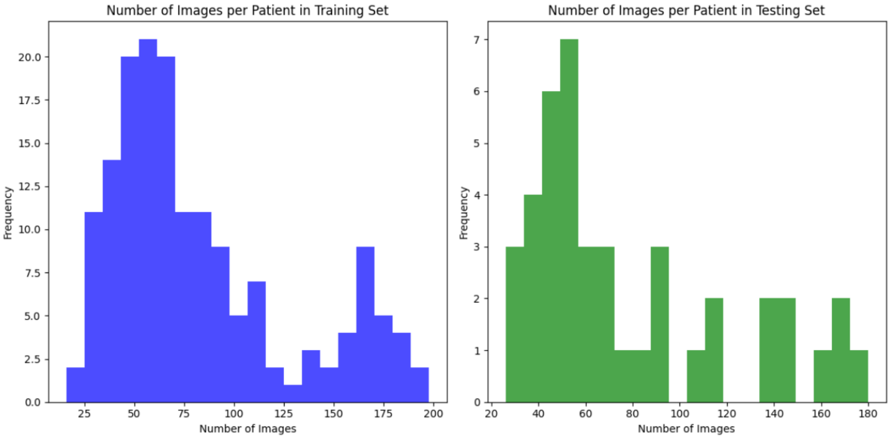

Figure 4: Number of images per patient in the train / test set