Deep Learning for Medical Imaging
MVA
Kaggle data challenge on Lymphocitosis classification

Simon Queric[*1,2] SIMON.QUERIC@TELECOM-PARIS.FR
Vincent Herfeld[*1,2] VINCENT.HERFELD@TELECOM-PARIS.FR
[1] *Télécom Paris*
[2] *ENS Paris-Saclay*

1. Introduction

Lymphocytosis is when a patient has a higher lymphocyte count than usual (or normal) and this can be due to several reasons. Some are not to be too worried about (reactive) since lymphocytosis can be caused by an infection or stress but another can be more worrying is lymphoproliferative disorder which is a type of cancer of the lymphocytes (tumoral). In today's common practice, diagnosis is done through visual microscopic examination of the blood cells joined by the study of clinical attributes such as age and lymphocyte count. This can help to estimate a malginancy type but this test is poorly reproducible. We must do additional tests that are efficient but time consuming and expensive : flow cytometry. We must choose wisely which patients will undergo this procedure. This is were our project comes in. We would like to automatise this selection process and in the most performing way. Our challenge is to build a machine learning model that is trained, thanks to accessible data, to learn how to efficiently classify patients that are either tumoral or reactive.

In this report we will share the solution we propose as well as how we chose our model and hyperparameters. We will share our various advances and results, which will be followed by rigorous validation to legetimize our solution.

2. Data

We quickly remind the data we have access to for this challenge. The available dataset is constituted of two modalities, the first are images of nucleated cells in blood smears produced then photographed automatically by hospital devices. The second is a dataset containing different patient attributes such as gender, date of birth, lymphocyte count and a label indicating a reactive case (0) or a tumoral case (1) for the training set and unkown case (-1) for the test set.

Blood smear images are relevant in a diagnostic context. Tumoral lymphocytes may have a higher nucleus-to-cytoplasm ratio and present abnormal morphology, such as irregular shape and increased size, compared to normal lymphocytes. Therefore, deep learning algorithms can be useful for classification tasks.

[*] Contributed equally

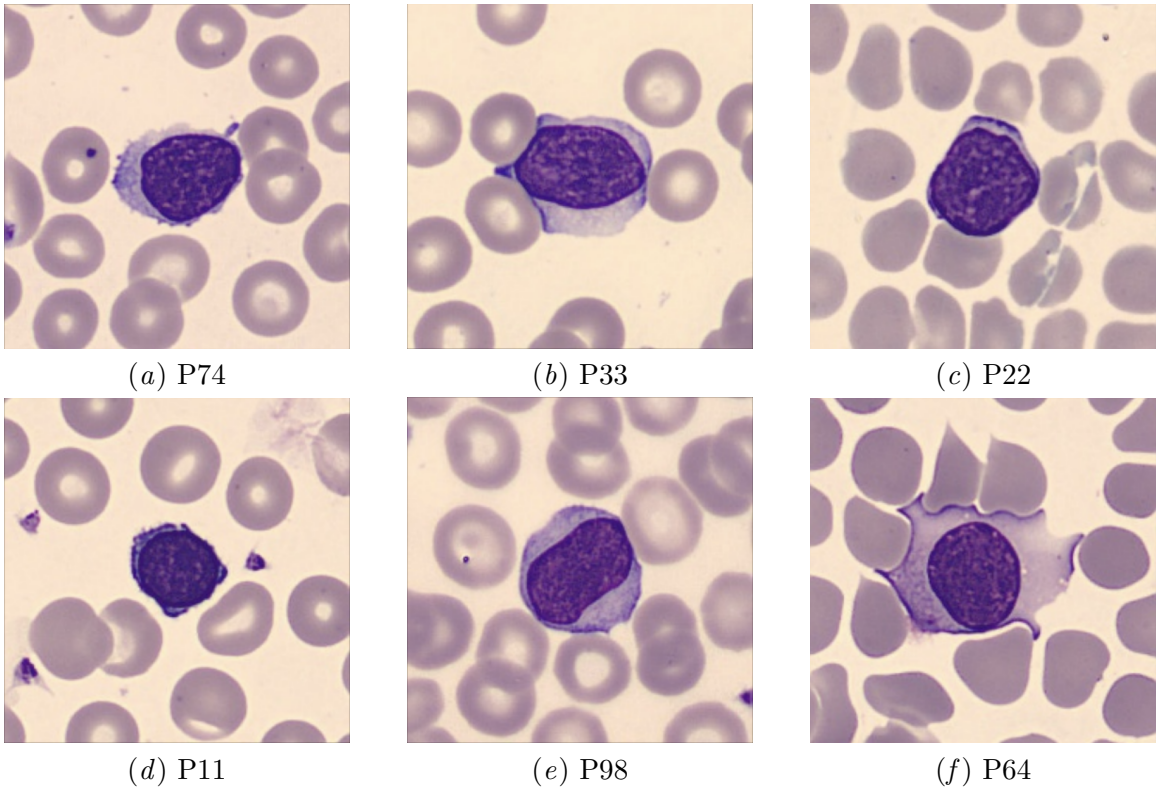
The training datas are unbalanced (142 subjects with 44 reactive and 98 malignant cases for training). Hence to evaluate our classification model we use the balanced accuracy score defined by $\text{BA} = \dfrac{\text{TPR} + \text{TNR}}{2}$.

Patients are mainly identifiable by their ID, this is usefull since their are several images per patient.

We mention that this data was provided by the work of Professor Pierre Sujobert and Universite de Lyon.

ID	LABEL	GENDER	DOB	LYMPH_COUNT
P46	1	M	1/7/1948	39.84
P45	0	M	4/24/1936	5.67
P101	0	F	19-12-1940	4.34
P150	1	M	3/18/1948	20.27
P116	1	F	9/26/1934	3.94

Table 1: Sample from the dataset table



(a) P74 (b) P33 (c) P22

(d) P11 (e) P98 (f) P64

Figure 1: First row corresponds to a lymphocyte cell from tumoral patients, second row are cells from reactive patients.

Normal cells should have a large cytoplasma area (as in P98 or P64), we can see that not all cells in reactive patients are normal as well as not all cells in tumoral patients are abnormal (P33). This presents a challenge for the problem we want to solve and can explain the difficulty for "manual" diagnosis from only a visual study (and in pratice restrained to only a few images) Showing every image to our model is a first to finding a solution.

3. Architecture and methodological components

Architecture Our main ideas are adapted from the architecture proposed by (et al., 2020). One of their methods is based on two classifiers : a multi-layer perceptron for the clinical data and a convolutional neural network which extract meaningful features from the images. They perform soft voting with the two classifiers. Our method is slightly different : we learn only one classifier. The inputs consist of patient blood smears images and their clinical data (age and lymphocyte count). The classifier comprises a feature extractor, followed by an average pooling of all the images features. To include patient clinical datas, we concatenated them to the previously obtained vector. The final layer is a fully connected classifier. The feature extractor chosen is a ResNet18 (He et al., 2015), as mentioned in (et al., 2020) and because it's a well known architecture to perform image classification.

At the end, for a collection of images $(X_j)_{1 \leqslant j \leqslant N}$, age and lymphocyte count of the patient (a, c), the probability of the patient having lymphocyte cancer is computed using $p = \sigma(f_\theta((X_j)_{1 \leqslant j \leqslant N}, a, c))$ where f_θ is the neural network, θ is the set of parameters and $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function. Our decision rule to predict the label is $\widehat{y} = \mathbb{1}(p \geqslant 0.5)$.

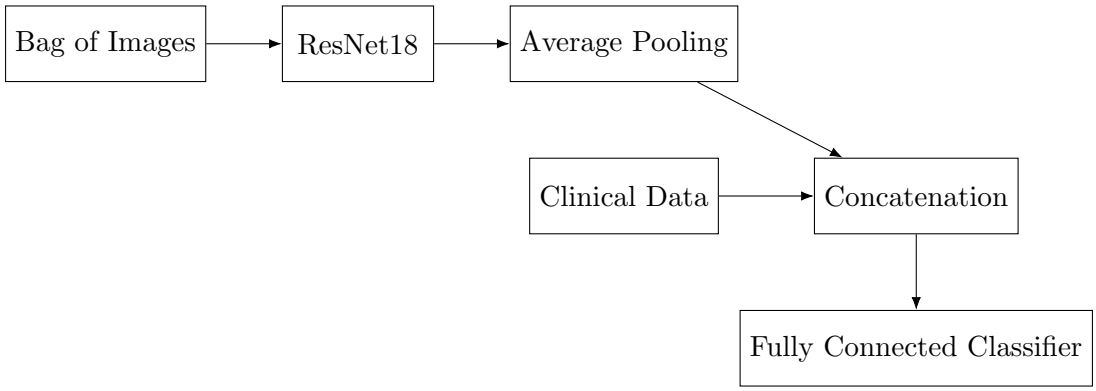


Figure 2: Our Pipeline for Lymphocitosis classification.

Preprocessing The only preprocessing step we applied is to clip the value of the images between 0 and 1.

Training The ResNet18 was trained from scratch over 20 epochs. Although (et al., 2020) notices more stability with a large batch size, we chose a batch size of 1. The loss function is the Binary Cross Entropy (BCE) loss. We use the PyTorch Deep Learning framework and the code is available at https://github.com/simonqueric/DLMI_challenge.

Validation To evaluate our method, we performed 5 cross-validation. It allows to avoid over-fitting.

4. Model tuning and comparison

This section discusses the ablation study, which involves modifying the learning rate, optimization method, and learning rate scheduler. Our study found that Stochastic Gradient Descent (SGD) resulted in greater stability than Adam. Empirically, SGD leads to flatter minima than the Adam optimizer.

Ablation study For the ablation study we train our network with different learning rates 10^{-3} and 10^{-4}, with a learning rate scheduler (exponential decay with decay $\gamma = 0.95$) or without a scheduler. We also test two optimisers : SGD with a momentum of 0.9 and Adam optimiser with its classical parameters. In the figure below we show the mean balanced accuracy on the validation set with half of the standard deviation (taken over all folds) in shaded colour. It's clear that our best model is the one trained with SGD, a learning rate of 10^{-4} and a decay of 0.95. We didn't experiment with alternative architectures like ResNet34 or EfficientNet because we observed during the initial epochs of training that these networks were over-parameterised for this task.

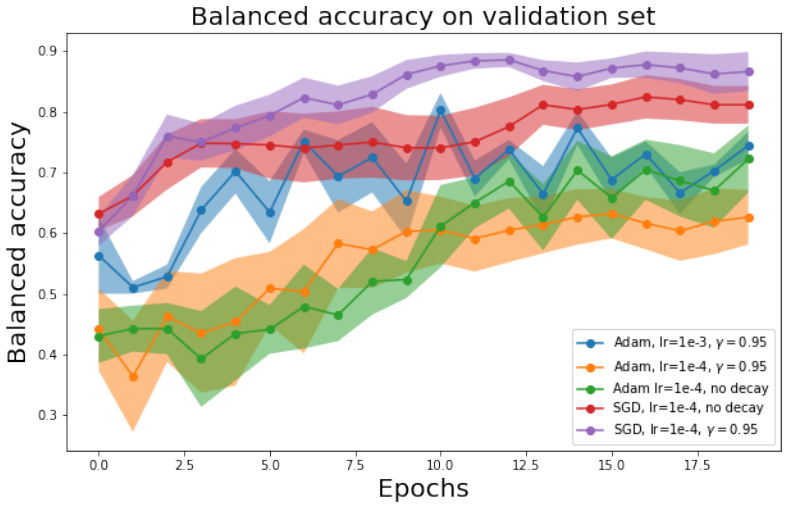


Figure 3: Ablation study

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
0.85	0.93	0.88	0.88	0.95

Figure 4: 5-fold cross validation, SGD, lr=10^{-4}, $\gamma = 0.95$

The table above describes the best balanced accuracy for the validation set on each fold. With a minimum balanced accuracy of 0.85, the method outperforms classical machine

learning methods, such as SVM or Random Forests, trained on the clinical datas only (according to the results reported in (et al., 2020)).

5. Conclusion and future work

We achieve a score of 0.83 (which doesn't depend on the fold we chose) on the Kaggle test dataset which is a lower accuracy than the one obtained with cross-validation.

Future work To validate the decision rule of our CNN, we should use Grad-Cam (Selvaraju et al., 2016) wich is a technique used in the field of computer vision to understand and visualize the regions of an image that are important for the prediction made by a CNN. With this technique, used in (et al., 2020), we can visualise if the CNN bases its decision on the morphology of the lymphocytes. Increasing the batch size could aslo be a way to improve our model as well as better image preprocessing or data augmentation.

References

Mihir Sahasrabudhe et al. Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis. *https://hal.science*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.