Classification of Lymphocytosis from Blood Cells

$Team - Léna\&Margaux$

Tornqvist Margaux Ezzine Léna

Abstract

This work focuses on predicting lymphocyte diagnosis for a patient given two types of data: lymphocyte microscopic images and clinical attributes such as age and lymphocyte concentration, by using deep learning methods. Convolutional neural networks are trained on bags of various size of microscopic images to extract meaningful features whereas a linear classifier is trained to predict the outcome from the clinical data. Moreover, we explore different ways of combining the outputs of each of these classifiers to predict the correct label for a given patient. We face the difficulties of multi-instance learning in a medical context, where even clinicians have variable consensus on the diagnosis. Finally our best results report a balanced accuracy of 78.96% *on the private leaderboard.*

1. Introduction

Lymphocytosis is a diagnosis characterized by lymphocyte count above 4.10^9/L. There are two forms of lymphocytosis: reactive and tumoral. The reactive form is most often due to an immune response to an infection or to stress. The tumoral form is due to leukemia, a lethal cancer attacking blood cells, and is therefore crucial to detect. Clinicians distinguish between the tumoral and reactive form by using their expertise on clinical data as well as images.

However this technique, prone to error, brings them to realize additional clinical tests. Notably, flow cytometry, is the gold standard to detect with high accuracy the correct form of lymphocytosis. But this expensive and time consuming technique can't be done for every patient. Considering a tool which could automatically classify lymphocyte population based on visual assessment and clinical data would save time and money to many hospitals and cliniques. This work focuses on exploring deep learning techniques which could be used to design a tool to automatically classify patients with high accuracy into reactive or tumoral form of lymphocytosis.

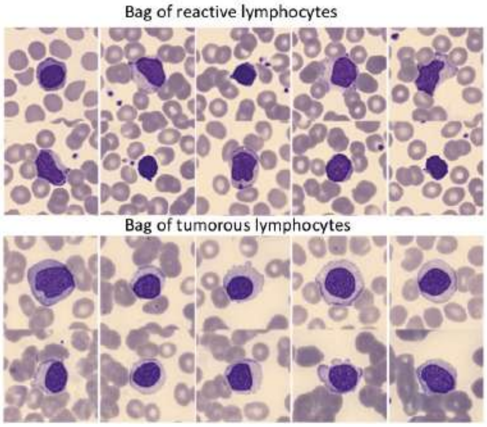

Figure 1. Two examples of bags of patients

2. Related works

Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis : This work recently published by Sahasrabudhe et al. was our first source of inspiration to conduct our work [2]. It relies on the same dataset as ours and tackles the same problem : developing deep learning methods able to predict if a lymphocytosis is reactive or tumoral. In their work, they use a CNN to extract features from images, and a MLP to predict the output from metadata. Moreover, their mixture-of-experts model combines information from these images as well as clinical attributes to form an end-to-end trainable pipeline for diagnosis of lymphocytosis. They report a balanced accuracy of 85.41% with their method.

Attention-based Deep Multiple Instance Learning : Multiple instance learning (MIL) is a variation of supervised learning where a single class label is assigned to a bag of instances. In this paper, the authors state the MIL problem as learning the Bernoulli distribution of the bag label which probability is fully parameterized by NNs [3]. Furthermore, they propose a NN-based permutation-invariant aggregation operator that corresponds to the attention mechanism. Notably, an application of the proposed attention-based operator provides insight into the contribution of each instance to the bag label.

3. Dataset

The dataset includes images and clinical data (age and lymphocyte count), from 204 patients of the hematology laboratory of LyonSud University Hospital. The centered and cropped images of lymphocytes were automatically extracted by an AI from larger microscopic images of blood smears. Each patient is associated to a bag of images, which size is correlated to the lymphocyte count, and a label, attributed by flow cytometry test. 142 subjects with 44 reactive and 98 malignant cases are used for training and 42 subjects with label to predict for testing. For exploring the different methods explained in the following section, we used a 80/20 split.

Label	Mean Age	Std Age	Mean LC	Var LC
0	55.9	20.5	4.9	0.9
1	75.7	11.6	31.9	44.3

Table 1. Training set, LC: Lymphocyte Count

Label	Mean Age	Std Age	Mean LC	Var LC
0	52.9	18.3	5.2	1.3
1	76.9	13.2	47.6	76.2

Table 2. Validation set, LC: Lymphocyte Count

In order to have a good representation of the patient population in both the training and validation set, as can be seen in the tables 1 and 2, we stratify the data on the diagnosis label and the age (by building age groups). Moreover, we are careful to randomly attribute each patient to either the validation or the training set. Otherwise the model may learn how to retrieve each image to each patient. We also tested building bags of fixed size, for example 30, by selecting them randomly from the bag of a patient in order to artificially increase the number of learning examples. This comes from an insight of Dr. Sujobert : Doctors never look at all the blood images, but only a sample of them is sufficient to diagnose whether the patient has tumorous cells or not. Images were resized to 224 x 244 and normalized. Classic data augmentation such as random rotations up to 20 degrees, vertical and horizontal flips were used to avoid overfitting and improve the robustness of our models.

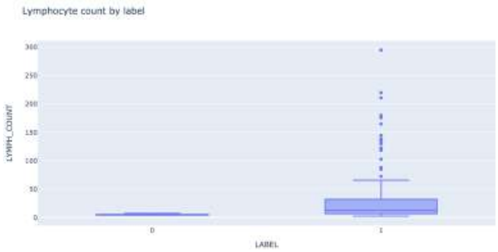

Figure 2. Lymphocyte count distribution per label

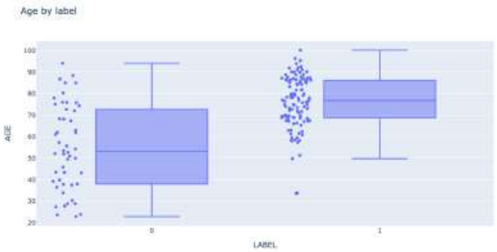

Figure 3. Age distribution per label

The two above figures show us that the lymph count and age are discriminative features to distinguish reactive patients from tumoral ones.

Now let us jump into the employed methods for our prediction task.

4. Methods

4.1. Understanding how doctors differentiate reactive and tumoral lymphocytosis

It is difficult for us, based on our personal visual assessment,to detect whether the lymphocytosis is tumorous or reactive. To get better insights, we called Pr. Pierre Sujobert, who provided the images for the data challenge. According to Pr. Sujobert, clinicians distinguish the two forms (reactive vs tumorous) first by analyzing clinical data such as lymphocyte count, age, concentration of white blood cells , blood platelets and absence or presence of particular cells in the blood. Then, they undertake a diagnosis from microscope lymphocyte pictures extracted from images of a blood smear from the patient. Clinicians pay attention in particular to:

- The size, color, quantity of ARNm in the nucleus,
- The size of the cytoplasm compared to the nucleus,
- The presence of granularities [1],
- The homogeneity of the bag of images: the diagnosis must be done for a population of lymphocytes and can't be done for a single image. Indeed, they try to assess whether the global lymphocyte population looks homogeneous, which correspond to a tumoral form, or heterogeneous, which is more characteristic of a reactive form.

The first 3 points are features that can be learned in the latent representation of a ResNet.

However, the last one, does not depend on a single image, but rather on the dependencies between all images of the bag. This is a keypoint that for now hasn't been exploited in previous multi-instance learning works : indeed, an underlying hypothesis is that all the images of a single bag are considered independent. Thus, the embeddings of each image are learned independently from the others in the latent space. There has been no attempt to learn inter-images representation, such as a homogeneity (or equivalently heterogeneity) score of a bag of images. Hence one question arises : how do we output, for each patient, a heterogeneity score for its bag of images ?

4.2. Notations

Let's introduce some notations useful for introducing our methods. We denote by $\{X_i^j\}$ the j-th image belonging to the bag of patient i, y_i^j it's instance level label and $y_i \in \{0, 1\}$ it's bag label. We introduce the embedding of each image belonging to a latent space of dimension E, and denote it $\{h_i^j\}$. Rather than considering that $y_i = max_j y_i^j$ we will consider a pooling function f_{pool} such that, at the instance level :

$$y_i = f_{pool}(\{y_i^j\}) \tag{1}$$

Or at a embedding level, followed afterwards by a linear classifier:

$$h_i = f_{pool}(\{h_i^j\}) \tag{2}$$

4.3. MLP on clinical data

In this section we train a simple feedforward neural network on the following metadata : the age of the patient, the lymphocyte count, and the number of images per patient. We use a 2-layer MLP with 100 hidden units and ReLU activation, followed by a linear layer with 2 outputs. We use the cross-entropy loss, and Adam optimiser set with a learning rate of 0.0001. This classifier undeniably overfits on the clinical data.

4.4. CNN on images

We train a classic CNN (pretrained on ImageNet), ResNet18, over 50 epochs, followed by a dropout layer of 0.25 and a linear classifier on the images to analyse how it learns only from visual data. In order to minimize the binary cross-entropy loss, we use Adam as optimizer with a learning rate of 10^{-4} and weight decay of 5.10^{-6}. We use as aggregation function, f_{pool} over the bag of images mean and max. Mean is definitely a better f_{pool}. The results aren't satisfying so we rather try to use the ResNet as a feature extractor to form an embedding $\{h_i^j\}$ for each image, and then apply a pooling function to form a unique embedding per bag. Passed through a linear classifier, this vector gives the final output y_i. We test different sizes of embeddings: 20, 30, 100. Increasing it don't seem to improve our results so we keep ourselves to a size of 20. We also test to train this model on bags of size 30 images and include weights to balance the data. As the validation loss increases gradually after the warm-up of the model, it seems that the model isn't learning anything. Table 1 shows that the sensitivity is close to 1 and the specificity close to 0, which is a proof that our model is naive and only outputs the most frequent label.

4.5. CNN + MLP

As the clinical data seems crucial to classify the patients we decided to use both the images and the clinical data. More precisely we combined the two previous cited methods. We averaged the outputs y_i^{CNN} and y_i^{MLP} of both classifiers (i.e $f_{pool} = f_{Mean}$. Moreover we tried to train a linear classifier upon these outputs by adding a fully connected layer upon the outputs to predict the final outcome (cf. figure 4).

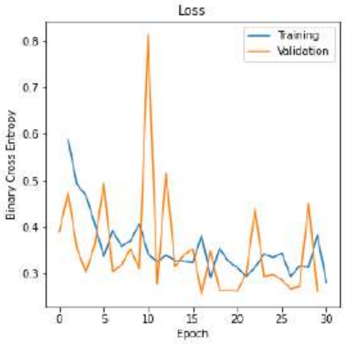

Figure 4. CNN+ MLP : Loss evolution for training and validation set over 50 epochs

4.6. Gated attention model + Embedding

This method inspired from the article of Ilse et al. [3], uses a Resnet-18 for the images, and applies the gated attention mechanism(GAM) to get a bag representation by using the weighted-attention sum as an aggregation function : $h_i = \sum_{j=1}^{n} w_i^j . h_i^j$. In our setting, the Resnet-18 is trainable. For the metadata (age, lymphcount, number of images) , we pass it through a linear layer to get a representation m_i, then we concatenate h_i and m_i to get g_i. Finally, we apply a classification linear layer on g_i to get 2 outputs, one for each class. We experiment with the cross-entropy Loss and the BCE Loss (in this case the final linear layer had only one output followed by a sigmoid). We use same training parameter settings as before. The figure below shows the evolution of the loss with epochs for train and validation set (cf. figure 5).

4.7. Gated attention model + MLP

Here we explore another method of combining the Gated Attention model outputs with the ones of the MLP trained on the clinical data. As for the previous methods, we try to average the outputs. The loss over the training on bags of size 30, is depicted in figure 5. Here we notice the role of the fixed size bags which seem to help the loss to converge. We

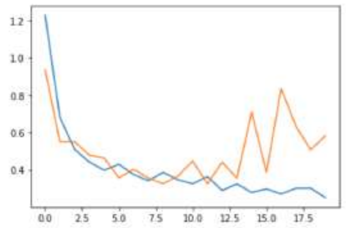

Figure 5. Gated Attention Model : Loss evolution for training and validation set over 20 epochs

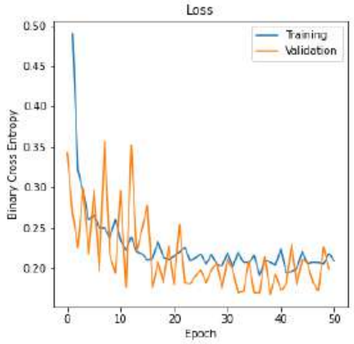

Figure 6. GAM + MLP : Loss evolution for training and validation set over 50 epochs

also try to train a linear regressor, denoted LR in the table of results which is actually a fully connected layer trained on the upon the concatenated outputs of the CNN and the MLP.

4.8. Heterogeneity scores for bags of images

When we called the doctor, he gave us an insight to detect patients with cancerous lymphocytes: if the lymphocyte images corresponding to a patient are too "similar", then the lymphocytes are tumoral. Now one question arises : how do we output a homogeneity (or heterogeneity) score for the bag of images of each patient ? There are sophisticated methods like Siamese neural networks, which require labels of pairs of images (1 if the 2 images are similar, 0 if not). But is there an unsupervised way to compute this score ? One straightforward way to do this would be to use the normalized variance of the flattened pixel vectors of the images.

$$Score = (\sum_{i=1}^{n} \sum_{j=i+1}^{n} (q_i - q_j)^2)^{0.5}/n \tag{3}$$

Where q_i is the flattened 1D vector of image i. We are aware that this is a naive method : indeed, if there are 2 images of the same lymphocyte, with one being the translation of the other, the variance wouldn't be 0, whereas it should be. Applying this method on our bags didn't give significant differences between the two classes of patients, as can be shown in the following table.

Heterogeneity score	mean	standard deviation
label 0	0.006527	0.000274
label 1	0.006541	0.000279

Table 3. Heterogeneity score mean and standard deviation for patients of class 0 and of class 1

Another method we tried is computing this normalized variance, but this time we replace q_i with the latent representation of the images in the last layer of ResNet18 (before the Fully-Connected layer). This is motivated by the fact that Resnet features are more meaningful and resistant to pixel biases like the one mentioned above. However there was, once again, not a significant difference.

Latent Heterogeneity score	mean	standard deviation
label 0	0.604862	0.066738
label 1	0.604987	0.073507

Table 4. Latent Heterogeneity score mean and standard deviation for patients of class 0 and of class 1

We think that this result is due to the fact that Resnet was not pretrained on medical images, hence it cannot leverage meaningful features from a very specific-domain images (here lymphocytes).

Another possible method that we didn't try for time reasons, is the following. Let M be the space of lymphocyte images. These images are distributed according to a probability distribution P on M. With this in mind, we could use a VAE on lymphocyte images, and use the obtained latent vectors as a new representation for the images, then compute the variance of these latent vectors for each patient. The advantage of this method is that it is unsupervised.

5. Results

Table 1 depicts our results for the different methods we explored. Let's recall that our models were selected on the validation set, and then retrained on the entire training before getting the predictions submitted to the kaggle challenge. First, we notice that training a CNN on the visual data is tricky. Introducing a pooling over the embeddings h_i^j within a bag seems to bring better results than averaging over the the direct outputs y_i^j. However, high sensitivity and low specificity for the trained CNNs shows their difficulty to learn the negative class. This is due to a high variability of the reactive bags of images which introduce instability in the training. To counter this, we though that training the model over bags of fixed size could bring stability and artificially increase the number of training examples.
Including the clinical data by adding the outputs of the MLP trained on the age and lymphocyte count increased

Algo	TestBalAcc	ValBalAcc	Loss	Spec	Sens
MLP	0.75584	0.8455	0.301	0.831	0.860
CNN $+f_{Mean}(\{y_i^j\})$	-	0.747	0.944	0.643	0.853
CNN $+f_{Mean}(\{h_i^j\})$	-	0.568	1.84	0.556	0.929
CNN$_{BAGS}$+f$_{Mean}(\{h_i^j\})$	-	0.967	0.207	0.793	0.953
CNN+MLP $+f_{Mean}(\{y_i^j\})$	0.78961	0.890	0.287	0.857	0.889
CNN+MLP $+f_{LR}(\{y_i^j\})$	-	0.968	0.091	0.969	0.986
GAM +Metadata Embed.	0.72252	0.9087	0.5831	0.9285	0.889
GAM$_{BAGS}$	0.77402	0.810	0.442	0.691	0.953
(GAM +MLP)$_{BAGS}$+f$_{Mean}(\{y_i^j\})$	0.63896	0.9328	0.179	0.759	0.981
GAM +MLP $+f_{LR}(\{h_i^j\})$	0.75584	0.979	0.068	0.971	0.989

Table 5. Evaluation of performance metrics for different models on validation set and test set if submitted on kaggle

the specificity. We also obtained the best balanced accuracy of 0.7691 on the test set with the $CNN + MLP + f_{Mean}(\{y_i^j\})$ model. However, the reactive cases seem still difficult to be learned. Moreover, as the MLP has tendency to overfit on the data, there is a great risk that the model is biased and won't perform as well on unseen data. Surprisingly, our gated attention model didn't perform as well than our CNN + MLP model even by combinating it with a MLP. Somehow, it is difficult to compare these models as they seem to be very sensible to the data which is noizy and includes high variability for the negative class. Finally, we may keep as best models our $CNN{+}MLP{+}f_{Mean}(\{y_i^j\})$ and the $GAM + MLP + f_{Mean}(\{y_i^j\})$, which seems to be our most robust model despite its lower balanced accuracy on the test set of 75.582%.

6. Interpretability

In this section we focus on the validation set and try to understand why our models have difficulties to classify the negative class and thus have such low specificity. For most of our models, the lymphocyte count of the false negatives is in the 25% inferior percentile of the histogram of lymphocyte counts for tumoral patients. Same goes for the age. Hence, given only the distribution of the age and the lymphocyte count for each label, it is statistically more significant that these patients have a reactive lymphocytosis.

Let's study more in details the validation results of our best model, the $CNN + MLP + f_{Mean}(\{y_i^j\})$. Let's recall that the y_{CNN} is obtained by averaging over the embeddings of all the images of a patient bag and that the final output corresponds to the average of y_{CNN} and y_{MLP}. Table 6 depicts the associated confusion matrix. Visualizing the false positive and negatives may help understand why our model struggles to classify certain patients.

	0	1
0	13	1
1	3	24

Table 6. Confusion matrix

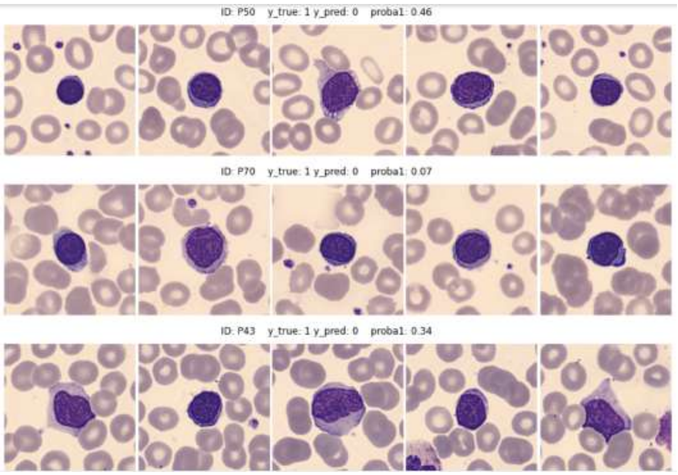
ID: P50 y_true: 1 y_pred: 0 proba1: 0.46
ID: P70 y_true: 1 y_pred: 0 proba1: 0.07
ID: P43 y_true: 1 y_pred: 0 proba1: 0.34

Figure 7. 5 images of each false negative

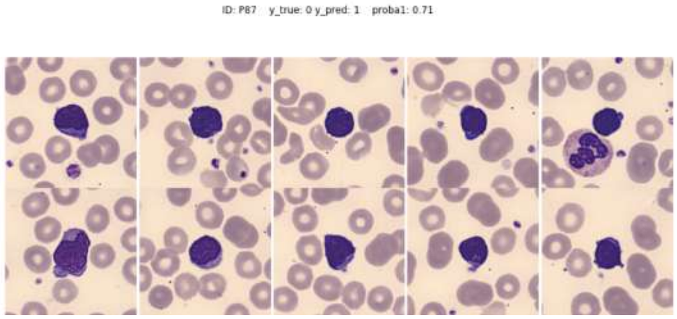

Figure 8. 10 images of each false positive

Somehow, it seems difficult to a non trained human eye to explain why the false negatives aren't seen as reactive by the network. As Figure 7 depicts, the two first cases may have been classified as reactive due to the small size and dark color of the lymphocytes. Somehow the third case seems difficult to explain. Indeed, this relies of the skills of a clinician used to do such diagnosis every day, knowing that even among clinicians there isn't always a consensus. Moreover, the unique false positive in Figure 8, doesn't seem to have high variability among the depicted lymphocytes. Perhaps, it is sufficient that one or two lymphocytes such as the one present on the top right hand image of Figure 8 ”pollute” the bag and make it think it is a tumoral label. A study of probabilities at an instance level could help understand such mechanisms.

7. Conclusion

To conclude, we noticed that despite the use of very sophisticated methods on the images like attention, our models have tendency to overfit on the metadata. Another difficulty is related to the small size of the dataset, which brings us ot work with a small validation set (41 images). Indeed, this explains instability in the validation loss during training. But what is even more surprising is that despite working with a stratified training and validation set, the validation set scores aren't representative of the test set scores.

Pr. Sujobert told us that one point which might help to make the right diagnosis about a patient is the concentration of other cells such as white blood cells and and cells present in the spinal cord. Hence, one way to improve the performances of the model is to use this additional metadata.

Moreover, one hypothesis of multi-instance learning is considering that the images of one bag are independent. In our case, and according clinicians, the homogeneity of a bag is a keypoint for tumor prediction. Hence, using more sophisticated methods that take into account inter-image representations, like Siamese neural networks to compute the homogeneity scores for the bags of images, may be very useful, both for enhancing predictions and for interpretability. However this requires to have annotated samples of pairs of similar and non similar images, which would require a doctor to annotate data.

References

[1] Hematology and clinical microscopy glossary. *College of American Histology*.

[2] E. Z. E. M. B. G. L. J. N. P. M. M. Sahasrabudhe, P. Sujobert. Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis. 2020.

[3] M. W. Maximilian Ilse, Jakub M. Tomczak. Attention-based deep multiple instance learning. 2018.