Classification of Lymphocytosis from Blood Cells
Final Project report

Aurélien Houdbert
CentraleSupélec
aurelien.houdbert@student.ecp.fr

Gladys Roch
CentraleSupélec
gladys.roch@supelec.fr

Céline You
CentraleSupélec
celine.you@student-cs.fr

Kaggle team : "Céline Gladys Aurélien"

March 2021

Abstract

Artificial intelligence is one of those technologies that makes the world safer, smarter and more convenient in various ways. It has been applied in several fields like security, law enforcement, marketing, health care industry, etc... Significant advances have been made in this last area. Thanks to deep learning and machine learning tools, it is already possible to track a patient's use of medication more accurately, detect genetic diseases with a high success rate and support pain management procedures. In this project, we will focus on blood cells and especially lymphocytosis. We will determine for patients whether the lymphocytosis is reactive (infection, acute stress, ...) or tumoral (manifestation of cancer of the lymphocytes). To tackle this classification task, we will implement a multiple instance learning algorithm.

1 Introduction

In typical machine learning problems like classification challenges, it is assumed that an instance clearly represents a class. However, in many real-life applications multiple instances are observed and only a general statement of the category is given. This is what we call multiple instance learning (MIL). MIL deals with a bag of instances for which a label is provided for the entire bag but not for individual instances within the bag. Hence, the main goal of multiple instance learning is to learn a model that predicts a bag label and to discover the key instances (the instances that trigger the bag label).

MIL is frequently encountered when dealing with diagnosis from medical images. Medical images often come in batches in order to encompass the 3D aspect of the body. MRI, CT are examples of medical exams generating bags of images.

We participated in a Kaggle competition on MIL for medical images, namely, "Classification of Lymphocytosis from Blood Cells". The objective of this competition is to propose the best approach to diagnose Lymphocytosis in patients using blood cells images and clinical data. Lymphocytosis is an increase in the number or proportion of lymphocytes in the blood and can be the manifestation of a lymphoproliferative disorder –a type of cancer. Diagnosis relies on visual microscopic examination of the blood cells together with the integration of clinical attributes such as age and lymphocyte count.

In this paper, we present our approach to solve the problematic of MIL for Lymphocytosis. We first detail the architecture of our model, then we present our numerical results and the impact of the parameters.

2 Related work

Deep learning and machine learning in the medical sector is not a novel subject. The medical sector produces a huge amount of imagery data that needs to be analysed. Lymphocyte images study is an important field of subject as blood is one of the best medical indicator in determining many pathological conditions.

For example, in [4] they adressed geometrical and statistical feature extraction from blood smear images. In [9] they focused on blood cell classification into subtypes using CNN approaches and in [6, 5], both papers implemented automatic detection and counting of lymphocytes from gastric cancer IHC images or immunohistochemistry images using deep learning techniques such as Faster R-CNN. Whereas in [11], the researchers studied spatial organization and molecular correlation of tumor-infiltrating lymphocytes based on HE images also using deep learning approachs.

Multiple instance learning has a variety of different applications and especially in the medical sector. Indeed, multiple instance learning (MIL) applies when the data is not labeled at the observation level but only labeled at a "bag" level (one label for multiple observations). Medical data can take a long time to label and can also be very expensive to produce as only experts in the domain can perform this labelling task.

In [3] they introduced a framework for weak supervision with convolutional networks by automatically selecting relevant images or regions within images from weak labels (attention mechanism).

A rich domain of application for MIL is histology. Indeed, histology easily generates huge amount of data (images of $200,000$ by $100,000$ pixel) that standard CNN can't process. The decomposition of these high dimensional images into smaller tiles forming bags of observations is mandatory in order to process it and MIL approachs are often required. In [1, 2, 8] they applied deep learning techniques to extract embeddings and attention scores for each tile of a bag.

The lymphocytosis diagnosis classification task was addressed in this paper [10] where they proposed a mixture of experts model mixing predictions of a convolutional network trainined on patients blood smears images and a multi layer perceptron trained on the patient's attributes such as age or lymphocyte count.

3 Data

The dataset for this challenge is made up of the data of 184 patients. There are 163 subjects with 50 reactive and 113 malignant cases for training and 42 subjects for testing.

Patient data include: a variable number of blood smears and patient attributes. The attributes are sex, date of brith, and lymphocytes count. Patients are between 23 and 100 years old and the lymphocytes count ranges between 2.28 and 295. The blood smears were automatically produced by a Sysmex automat tool, and the nucleated cells were automatically photographed with a DM-96 device. There are between 16 and 198 images per patient. Figure 1 shows examples of normal and abnormal lymphocytes.

The training and test datasets are balanced in terms of data distributions. The following tables show statistics on each set and on the global pool of patients. Both sets include approximately half women

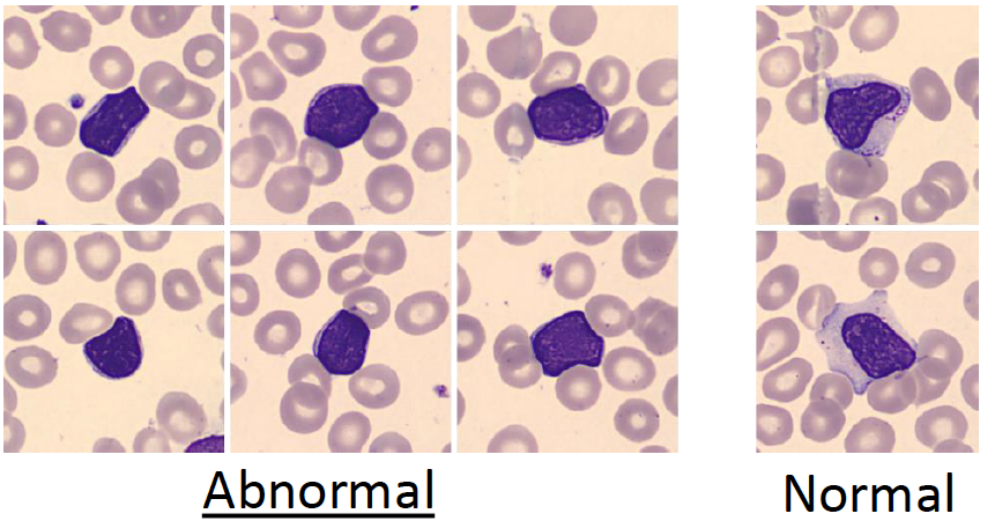

Figure 1: Normal and abnormal lymphocytes samples

and half men. They both present a similar age and lymphocytes count distribution, however the training set is not perfectly balanced with a 69% ratio between classes.

	LABEL	GENDER	LYMPH_COUNT	age
count	163.00	163.00	163.00	163.00
mean	0.69	0.51	26.42	69.76
std	0.46	0.50	46.78	17.66
min	0.00	0.00	2.28	23.00
25%	0.00	0.00	5.04	61.00
50%	1.00	1.00	7.81	73.00
75%	1.00	1.00	20.44	84.50
max	1.00	1.00	295.00	100.00

Figure 2: Description of the training set

From the data, we were able to detect that sick patients have a higher lymphocytes count (about 7 times more than healthy patients) and we also noticed that patients suffering from tumoral lymphocytosis are quite old and tend to be more men than women.

	LABEL	GENDER	LYMPH_COUNT	age
count	42.0	42.00	42.00	42.00
mean	-1.0	0.57	24.37	66.69
std	0.0	0.50	44.11	19.73
min	-1.0	0.00	4.08	22.00
25%	-1.0	0.00	5.17	56.00
50%	-1.0	1.00	6.76	68.00
75%	-1.0	1.00	22.38	82.50
max	-1.0	1.00	217.59	98.00

Figure 3: Description of the test set

4 Architecture and methodological components

The architecture we chose to implement combines two learning pipelines: a Multi Layer Perceptron (MLP) that learns from the clinical attributes of the patients (age and lymphocytes count) and a MIL classifer that learns from the blood smears images. To combine these two pipelines we chose to follow the work of Sahasrabudhe and al. [10]. In their paper they proposed a Mixture of expert (MOE) model. The MOE improves on a simple average-voting ensemble classifer, by learning the weights to associate to each model (MLP and MIL). This more-advanced way of combining two learning pipelines has proven to increase the balanced accuracy from a classical approach.

Figure 4 illustrate the pipeline we implemented.

4.1 Multi-layer Perceptron for Clinical Data

Multi-layer Perceptrons offer state-of-the-art results in classification task. We used a simple 3-layer architecture (3 fully connected layers *2@128, 128@256, 256@1* which accounts for 33,665 trainable parameters) to classify patients using only their clinical attributes: age and lymphocytes count. The output prediction of the MLP is noted $\hat{y}_{\text{MLP}}$. The paper [10] originally used a much more simple MLP model. Indeed their model was only composed of 2 fully connected layers (*2@2, 2@1*) but after training over 100 epochs, the loss plateaued rapidly and the accuracy and balanced accuracy recorded on the validation set showed that the model outputs almost random predictions (balanced accuracy of 0.5).

We confirm the choice of MLP by comparing its performances with classical machine learning algorithm: see Table 4.1.

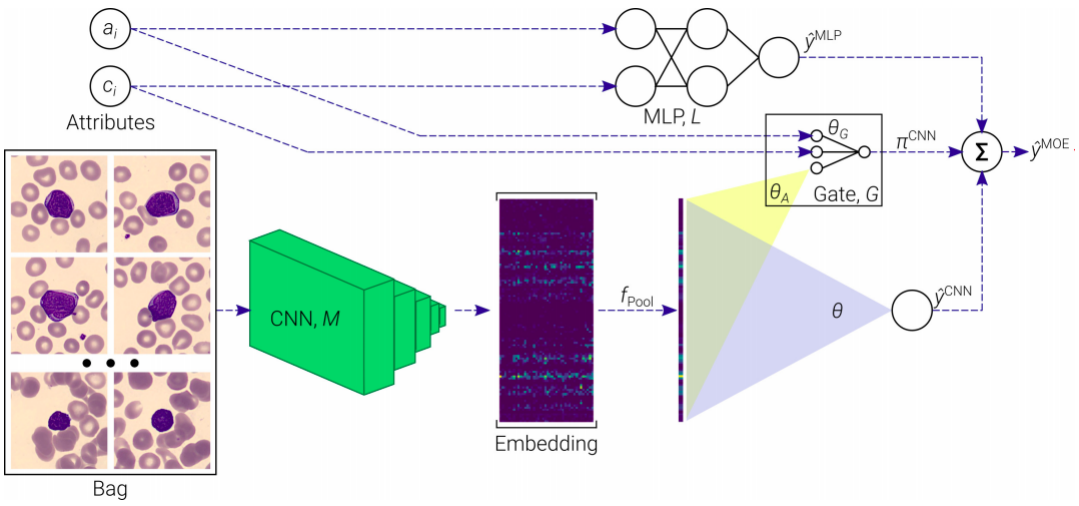

Figure 4: Mixture Of Expert (MOE) model architecture from [10]

Model	Accuracy	Balanced Accuracy	Recall/Sensitivity	Precision/Specificity
SVC	0.76	0.74	0.81	0.81
SGD Classifier	0.82	0.86	0.71	**1.0**
Naive Bayes (gaussian)	0.85	0.88	0.76	**1.0**
Random Forest	0.79	0.78	0.81	0.85
MLP	**0.91**	**0.89**	**0.95**	0.91

Table 1: Comparison of MLP classifier with other machine learning classifiers. Validation metrics - 33 patients (train size=0.8, random sate=2)

4.2 Multiple Instance Learning (MIL) for blood smears

In order to classify bags of images, we first produce an embedding of the bag using a Residual Network (ResNet) approach, this embedding is then passed into a linear classifier.

The ResNet [7] is a Convolutional Neural Network (CNN) with identity skip connections preventing vanishing gradient. Our ResNet has 4 successive convolutions blocks and outputs a 25,088-dimensions embedding of the input. To build the CNN network, we tried different dimensions (in $\{32, 64°\}$ for the residuals of the resnet. The main trouble we ran into when increasing the dimension of the residuals, was the memory. Indeed we trained our model on google colab's GPU and we were quite limited in memory. This is the reason why we integrated the dataloader for each patient directly inside the model. Each image of the bag of blood smears is passed in parallel through this network, the outputs are averaged together in order to obtain a single embedding for the whole bag.

To increase the coherence of the bag embedding we tried implementing an attention-based combination of the embeddings following this paper's [8] implementation. Indeed, the intuition behind attention is that blood smears of a same patient might not all carry the same information and don't have the same importance for the diagnosis. Instead of simply averaging the embeddings, the attention gate is a way to put different weights on different blood smear. This way, images' embeddings with

meaningful features might be given larger weights and therefore resulting in better predictions.

The final embedding of the bad of images is passed through a one-layer linear classifier that outputs a score which gives (when passed through a sigmoid) the probability of the patient being sick, $\hat{y}_{\text{CNN}}$.

4.3 Mixture of Expert (MOE)

The classical approach of combining two networks learning in parallel is to average their respective predictions to obtain the final prediction. The Mixture Of Expert (MOE) architecture aims at learning the best way to combine the two models. The MOE is a network that takes as input the inputs of all the classifiers involved in the global model. In our case the two classifiers are the MLP and the linear classifier of embedded images. Therefore, the input of the MOE is the tuple made up of the clinical attributes and the embedding of the bag of images. This architecture can be seen in Figure 4.

The MOE does not output a final prediction for the patient diagnosis but rather the weight to give to the CNN classifier, π_{CNN}, in the average of the two models. The weight given to the MLP is then $1 - \pi_{\text{CNN}}$. The final diagnosis is

$$\hat{y}_{\text{MOE}} = \pi_{\text{CNN}} \times \sigma(\hat{y}_{\text{CNN}}) + (1 - \pi_{\text{CNN}})\, \sigma(\hat{y}_{\text{MLP}})$$

where σ is the sigmoid function.

There are multiple ways of training a MOE. We chose to first train the CNN on blood smears (embedding + classification) and the MLP on attributes, independently and then to train the MOE using the pretrained networks. The reason we chose this training pipeline is that it is simpler than learning the MLP and the CNN along with the MOE, it is also the approach giving the best results in [10]. The downside of this approach is that the embedding learned for the bags of images do not leverage the knowledge hold by the attributes. On the other hand, it offers more visibility on the performance of the individual models and allow for easier fine tunning.

5 Training and Results

5.1 Evaluation metric

Because our dataset is a bit imbalanced, we evaluated the quality of our model using the *balanced accuracy* metric that takes into account the potential imbalance between classes in the data.

Before defining the balanced accuracy, lets define the Sensitivity and the Specificity. The Sensitivity measures the proportion of *positives* that are correctly identified and the Specificity measures the proportion of *negatives* that are correctly identified.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$
$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where *TP* are the *true positives*, *TN* are the *true negatives*, *FP* are the *false positives*, and *FN* are the *false negatives*.

The balanced accuracy is simply the average of the sensitivity and the specificity:

$$\text{Balanced accuracy} = \frac{Sensitivity + Specificity}{2}$$

5.2 Training process

In order to train the models that we defined in the previous sections, we first split our dataset into a training set (80%, 130 patients) and a validation set (20%, 33 remaining patients) to record the performance of our model during the training and prevent overfitting.

As a first experiment and in order to establish a personal baseline, we built a first classic machine learning model (SVM) using only the patient attributes (age and lymphocyte count). This initial baseline obtained a balanced accuracy score of 0.75 on the kaggle competition.

We implemented several architectures and submitted the results on the kaggle competition to evaluate the performance. We first tried a simple MIL model composed of a ResNet coupled with a fully connected classifier that took as input the averaged embeddings given by the ResNet. We managed to reach a balanced accuracy score of 0.78 on the Kaggle test set. With the idea that combining embeddings in a smarter way than a simple average could improve our results, we implemented an attention gate to learn how to perform a weighted sum of the embeddings. We obtained very good performances on the validation set but results didn't scale when submitted on the kaggle competition.

Following [10], we introduced more complex models as described in section 4. We trained both the MIL and the MLP for 100 and 150 epochs respectively using Adam optimizer with a learning rate of 10^{-4}. The MOE was then trained on top of these two frozen model for 200 epochs using Adam optimizer with a learning rate of 10^{-4}. The training loss of the three trainings are shown in figure . To make sure that we were not overfitting the training data in any time, we performed early stopping by recording the loss on the validation set at each epoch and saved the model when the validation loss was at its lowest.

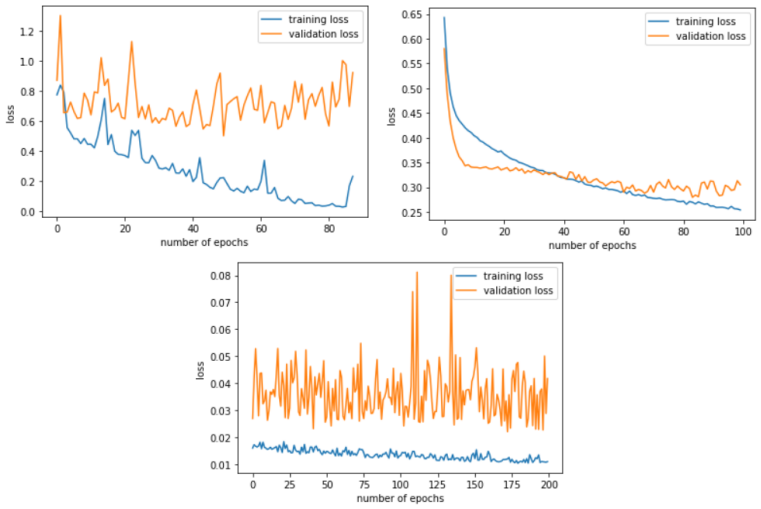

Figure 5: CNN, MLP, MOE training losses. **Top-left** CNN training loss and validation loss, **Top-right** MLP training loss and validation loss, **Bottom** MOE training loss and validation loss

5.3 Data augmentation

The training set contains the data of 163 patients each having between ≈ 20 and ≈ 190 blood smears. This dataset is rather small and the complexity of the designed model makes the training difficult and especially because of rapid overfitting. To tackle this issue we introduced data augmentation. We choose to add random horizontal and vertical flips as well as rotations from the set $\{0, 90, 180, 270\}$. Incorporating data augmentation is a way to artificially increase the dataset size and therefore reduce overfitting.

5.4 Results

Surprisingly, the model that obtained the best score on the kaggle competition is the MLP model alone and reached a balanced accuracy score of **0.865**. Our mixture of experts model composed of the MIL and the MLP obtained a balanced accuracy score of **0.797** on the kaggle competition. During the training process, when looking at the results on the validation set, the MOE model seemed to outperform both the MLP and MIL by combining predictions of the two models (Figure).

```
MOE Val Accuracy Score ->  0.9394 ; Val Balanced Accuracy Score ->  0.9345
CNN Val Accuracy Score ->  0.7576 ; Val Balanced Accuracy Score ->  0.7202
MLP Val Accuracy Score ->  0.8788 ; Val Balanced Accuracy Score ->  0.869
```

Figure 6: Training results of MOE training with early stopping on the validation loss

The training losses shown in figure led to very good results on the validation set. We were surprised about the testing results because the MOE gave very impressive results on the validation set (table) but performed "poorly" on the testing set (0.797 balanced accuracy). One possible reason for that could be the difference balance between positive cases and negative cases in the validation set and testing set although we tried to create a validation set as representative as possible. Another explanation could be that the MOE was "lucky" on the validation set as the validation loss curve could imply. Indeed, looking at this curve, there is no evidence that the MOE actually learned anything during the training.

The results we obtained for this challenge are summarized in the table below:

AVG	MOE	MLP
0.748	0.797	**0.865**

- **AVG**: CNN and MLP outputs are averaged to produce a final prediction
- **MOE**: Results given by the MOE taking as inputs the CNN and MLP scores
- **MLP**: Mutlilayer perceptron alone

6 Conclusion

During this project we implemented different machine learning and deep learning network architectures. We first obtained a reasonable balanced accuracy score of 0.75 using a simple SVM applied on the patient attributes age and lymphocyte count. This result was improved to 0.79 using a mixture of

experts model based on a ResNet + classifier extracting embeddings and predictions from blood smears and a multilayer perceptron applied on the patient attributes. However, the best overall model was in the end a well trained MLP alone that reached a balanced accuracy score of **0.87**.

References

[1] Pierre Courtiol et al. "Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach." In: (2018). URL: https://arxiv.org/abs/1802.02212.

[2] Olivier Dehaene et al. "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology." In: (2020). URL: https://arxiv.org/abs/2012.03583.

[3] Thibaut Durand, Nicolas Thome, and Matthieu Cord. "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks." In: (2016). URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Durand_WELDON_Weakly_Supervised_CVPR_2016_paper.pdf.

[4] Abdullah Elen and M. Kamil Turan. "Classifying White Blood Cells Using Machine Learning Algorithms." In: *International Journal of Engineering Research and Development* (2019).

[5] I. Keren Evangeline et al. "Automatic Detection and Counting of Lymphocytes from Immunohistochemistry Cancer Images Using Deep Learning." In: *Journal of Medical and Biological Engineering* (2020).

[6] Emilio Garcia et al. "Automatic Lymphocyte Detection on Gastric Cancer IHC Images using Deep Learning." In: (2017). URL: https://ieeexplore.ieee.org/document/8104187.

[7] Kaiming He et al. "Deep Residual Learning for Image Recognition." In: (2015). URL: https://arxiv.org/abs/1512.03385.

[8] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. "Attention-based Deep Multiple Instance Learning." In: (2018). URL: https://arxiv.org/abs/1802.04712.

[9] Tiwari Prayag et al. "Detection of Subtype Blood Cells using Deep Learning." In: *Cognitive Systems Research* (2018).

[10] Mihir Sahasrabudhe et al. "Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis." In: (2020). URL: https://hal.archives-ouvertes.fr/hal-03032875.

[11] Joel Saltz et al. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images." In: *Cell Reports* 23 (2018), pp. 181–193.