Deep Learning for Medical Imaging challenge
Lymphocytosis classification

Ben Kabongo ben.kabongo_buzangu@ens-paris-saclay.fr
Junior Cedric Tonga juniortonga2022@gmail.com
Master MVA, ENS Paris-Saclay

March 2024

1 Introdutcion

As part of the Deep Learning for Medical Imaging course, we took part in the **Lymphocytosis classification** challenge. Lymphocytosis, characterized by an elevated lymphocyte count, indicates an above-average presence of lymphocytes in the bloodstream. Lymphocytes, a subset of white blood cells, are pivotal components of the immune system, aiding in the defense against infections by combatting pathogens. We have at our disposal a training dataset containing various information on patients, as well as a bank of images for each patient. If we consider that a single image of a lymphocyte can indicate the presence of cancer in a patient, then we can frame the problem as one of **Multiple Instance Learning (MIL)** [1]. The aim of the challenge is to propose models for binary classification, indicating whether each patient is reactive or malignant. Our implementation is available here: https://github.com/BenKabongo25/mva-dlmi-lymphocytosis-classification-project.

2 Data exploration

To build up the dataset, blood smears and some attributes of 205 patients were collected and then anonymized. We have 163 subjects for training, including 50 reactive and 113 malignant cases, and 42 subjects for testing. Each subject is assigned an identifier, a label (0 reactive, 1 malignant, -1 for test subjects), a gender (M or F), a date of birth (which we later convert into an age for better processing in our models), a lymphocyte count and finally a bank of between 20 and 180 images of bloom smears. There are almost as many male as female subjects. For training data, subject ages range from 26 to 103, with a mean of 72.76 and a median of 76. Lymphocyte count ranged from 2.28 to 295, with a mean of 26.42 and a median of 7.81. Table 1 presents some statistics as a function of label, for age and lymphocyte count respectively.

Attribute	Label	Count	Mean	Std	Min	25%	50%	75%	Max
Age	0	50.0	58.26	19.77	26.0	41.5	56.5	75.75	97.0
	1	113.0	79.18	12.04	37.0	72.0	80.0	89.00	103.0
Lymphocyte count	0	50.0	5.0052	1.0032	4.01	4.2225	4.505	5.7075	7.68
	1	113.0	35.9012	53.5728	2.28	6.94	12.43	31.11	295.00

Table 1: Statistics on the age and lymphocyte count of training set subjects.

3 Methodology

3.1 Metrics

As the class distribution of the training data is unbalanced, the main metric for evaluating our models is balanced accuracy.

$$\text{balanced accuracy} = \frac{\frac{TP}{TP+TN} + \frac{TN}{TP+TN}}{2}$$

with TP the number of true positive samples and TN the number of true negative samples. We also evaluate our models on accuracy and f1-score.

3.2 Models

With attributes and image banks for each patient, a natural first idea is to train two models, a first model on attributes and a second model on images. And then to combine the predictions of the two models with an aggregation operator, such as a weighted average, or a decision rule. In the following, we present the different models we have considered, for attributes and images respectively, as well as the different decision rules used for prediction.

3.2.1 Attribute-based models

The attributes associated with each patient are an identifier, date of birth, gender and lymphocyte count. For the models we have considered, we exclude the patient identifier; and we transform the date of birth into age. For some models, we use the onehot encoding for gender.

Linear models. The linear models we considered were logistic regression (LR) [11] and support vector machine (SVM) [12]. These are binary classification models (which can be extended to multiclass). Logistic regression reduces the classification problem to a binomial regression problem; support vectors treat the classification problem as a quadratic optimization problem.

Decision tree and random forest. Decision trees [2] represent a non-parametric approach in supervised learning, applied for both classification and regression tasks. The objective involves constructing a model that forecasts the target variable's value through the acquisition of straightforward decision rules derived from the features within the data. A random forest is a meta-estimator that builds multiple decision tree classifiers on different subsets of the dataset and uses averaging to improve prediction accuracy while reducing overfitting.

Neural network. We also used neural network models based on patient attributes to solve the binary classification task. For model optimization we used the Adam optimizer [6] with cross-entropy loss.

Model	Hyper-parameter	Values						
Logistic regression	C	10^{-6}*	10^{-3}	10^{-2}	10^{-1}	1	10	100
	Solver	lbfgs*	liblinear	newton-cg*	sag	saga		
	Penalty	None*	L1	L2				
	L1 ratio	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	
SVM	C	10^{-6}	10^{-3}	10^{-2}	10^{-1}	1*	10	100
	Kernel	Linear	Poly	RBF*	Sigmoid			
	Degree	2	3					
	Gamma	Scale*	Auto					
Decision Tree	Criterion	Gini	Entropy	Log loss*				
	Splitter	Best*	Random					
Random Forest	Criterion	Gini*	Entropy	Log loss				
	Num. estimators	5	10	20*	30	50	100	200
Neural Network	Optimizer	Adam						
	Learning rate	10^{-1}	10^{-2}*	10^{-3}				

Table 2: Hyper-parameters considered for each attribute-based model. For each model, the value of each hyper-parameter giving the best balanced accuracy of the model is marked with a star.

We use the scikit-learn library to implement linear models, decision trees, and random forests, and the PyTorch library to implement neural networks. We have therefore referred to the documentation for the various models to list the relevant hyper-parameters. For each model, we perform a grid-search on the hyper-parameters to find those that maximize balanced accuracy. Table 2 lists the hyper-parameters considered for each model.

3.2.2 Image-based models

In addition to patient information, each patient has a number of images. These images enable us to design image-based models. The number of images per patient varies. However, the size of the original images is the same: 224 x 224 RGB images. Different classes of models and approches are considered in image-based models: one-to-one models, many-to-one models and multiple instance learning approch.

One-to-one models. These are models that consider each patient image individually and therefore predict the image class. The expected image class is the class associated with the patient. To balance the number of images in the two classes, we perform data augmentation in some cases.

Many-to-one models. These models consider several images of a patient together, and therefore predict the class of a bag of images. The expected class of the bag of images is the class associated with the patient. For these models, the size of the image bag is fixed; it may happen that several disjoint bags are built for certain patients, so in some cases there may be several predictions to take into account.

For both classes of models, we consider convolutional neural networks (CNNs) [7] as well as pre-trained models such as VGG16 [10], ResNet18 [3] or ResNet50 [3] to obtain the features of an image. In the case of pre-trained feature extractor (VGG16, ResNet), we freeze the extractor; the extracted features are then considered as input to aggregation and classification models. In a situation where the model returns several predictions (image by image or by bags of images), we use the following aggregation operators: the majority class, the minimum (class 0) or the maximum (class 1). We remove the classification head from the pre-trained models. For one-to-one models we add a binary classification head on top of the feature extractor under consideration, and for many-to-one models we consider different feature aggregation processes. We use various image feature aggregation functions for many-to-one models to create a single feature set. Once the feature has been obtained by the aggregation operator, we use a multilayer perceptron for the classification task.
The aggregation functions considered are: **simple aggregation** like the minimum, maximum or sum per dimension of each image's features and **attention** [13]. For attention, the input to the model is the set of image features extracted from the image bag. The features are then aggregated by weighting the feature of each image using the attention mechanism.

MIL model. As mentioned in the introduction, this problem can be classified as a Multiple Instance Learning (MIL) problem, which is a generalization of binary supervised learning problems where training class labels are associated with sets of bags rather than individual instances. Regarding the methodology of MIL [9], a bag is considered negative if all its instances are negative, and a bag is considered positive if at least one (or a small ratio) instance is positive. In the context of our problem, positive patients must contain at least one instance classified as positive (images containing information indicative of cancer detection), while negative patients must have all their instances classified as negative (images not relevant to cancer detection). This can be implemented using instance predictions or instance embeddings.

In our Multiple Instance Learning (MIL) approach, patients represent the bags and their corresponding images represent the instances. The data loader is configured to return images along with their labels, bag identifiers, and indices, allowing us to organize the images belonging to each patient. For a given batch of patients, we gather all the images and stack them into a batch. This batch is then fed independently into the backbone network, without applying global pooling. Since the number of images per patient varies and we pass a large quantity of images to the backbone, we only retained a random subset of images on which gradients will be propagated. Next, we apply a 1x1 convolution with a single output channel to the resulting feature maps, generating pixel-wise prediction logits. These logits are then rearranged into bags of scores, where each bag corresponds to all predictions made for one patient. Finally, a **custom aggregation** processes each bag of scores to produce the final prediction for each patient.

Building upon the approach described in this article [8], our aggregation operator tends to adjusts weights to prioritize gradients for predictions considered crucial, particularly in scenarios with imbalanced classes. Given all the logits (prediction scores) $l_i^j, i = 1, \ldots, n_j$ for a patient j; inspired by [4], for aggregating (called **custom aggregation** in the results table 5) these logits we introduced weights $a_i^j \geq 0$ (sum all to 1) such that the final logit for patient j is given by $l^j = \sum_{i=1}^{n_j} a_i^j l_i^j$.

The weights a_i^j are defined by applying batch normalization [5] to the logits while stopping the gradient followed by the sigmoid function : $a_i^j = \sigma\left(\beta + \gamma \frac{l_i^j - \mu}{\sqrt{\sigma^2 + \epsilon}}\right)$, where γ and β are trainable parameters, ϵ is a small constant, and μ and σ^2 are respectively the mean and variance of the logits l_i^j over a batch of patients and n_j the number of images of the patient. To obtain the predicted class, we simply apply the sigmoid function to the final logit to obtain the probability score y_j ($y_j = \sigma(l^j)$). Then, we compare this probability to 0.5: if the probability is greater than 0.5, we assign the class label 1, and if it is less than or equal to 0.5, we assign the class label 0.

3.2.3 Decision rules

Attribute-based and image-based models can be trained separately, providing a prediction of each patient's class. Let $\hat{y}_a$ denote the prediction of a model based on attributes and $\hat{y}_i$ the prediction of a model based on images. We used the following decision rules to combine the predictions from the different models:

- α-**Weighted sum**: $\hat{y} = 1$ if $\alpha \hat{y}_a + (1 - \alpha)\hat{y}_i > 0.5$ else 0, with $\alpha \in [0, 1]$.

- **Maximum**: $\hat{y} = \max(\hat{y}_a, \hat{y}_i) \Leftrightarrow \hat{y} = 1$ if $\hat{y}_a = 1$ or $\hat{y}_i = 1$ else 0.
- **Minimum**: $\hat{y} = \min(\hat{y}_a, \hat{y}_i) \Leftrightarrow \hat{y} = 0$ if $\hat{y}_a = 0$ or $\hat{y}_i = 0$ else 1.

4 Experiments and results

Model	Accuracy	Bal. accuracy	Leaderboard
Decision Tree	80.5	79.5	-
Random Forest	87.8	88.2	78.961
SVM	82.9	89.1	72.207
Logistic regression	82.9	89.1	83.116
Neural Network	**90.0**	**90.0**	**85.714**

Table 3: Attribute-based models accuracies, balanced accuracies and balanced leader accuracies on validation set.

Feature extractor	Predictions aggregator	Accuracy	Balanced accuracy
CNN	Min	48.78	65.00
	Max	73.17	50.00
	Mean	75.61	54.54
CNN Augmented data	Min	58.00	60.50
	Max/Mean	42.00	50.00
VGG16	Min/Max/Mean	65.85	50.00
ResNet18	Min	60.97	63.23
	Max/Mean	58.53	50.00
ResNet50	Min	31.70	51.72
	Max/Mean	70.73	50.00

Table 4: One-to-one image-based models accuracies and balanced accuracies on validation set for different predictions aggregators

Feature extractor	Features aggregator	Accuracy	Balanced accuracy	Leaderboard
CNN	Max/Attention	70.00	50.00	-
ResNet18	MIL custom aggregation	-	86.00	82.337
ResNet34	MIL custom aggregation	-	**88.16**	**85.714**

Table 5: Best Results of Different Backbone and Aggregation Methods

Attribute-based models. For each attribute-based model, we listed a few hyper-parameters that we felt were relevant, and for each of these hyper-parameters we made a choice of a list of values (see Table 2). For all models, we used 75% of the data for training and 25% for validation. Then, using a grid search, we looked for the combination of hyper-parameters that maximized the balanced accuracy score for each model. Outperforming the linear and decision tree models, the neural network model achieves the best performance, with a balanced accuracy of **90% locally** and **85.714% on the leaderboard**, as shown in the Table 3 of attribute-based model results.

Image-based models. We have considered different feature extraction models: CNN, VGG16, ResNet. For one-to-one models, we tested different prediction aggregation functions. One-to-one models fail to achieve satisfactory balanced accuracy. The results reported in the Table 4 show that the best one-to-one model in terms of balanced accuracy is the model with CNN for feature extraction and the minimum function for prediction aggregation. This model has a balanced accuracy of 65%. We note, however, that the accuracy of this model remains very low compared to other one-to-one models. As for many-to-one models, they also fail to achieve better balanced accuracy than attribute-based models. The Multiple Instance Learning approach achieves better performance.

MIL model. For this model, We experimented with two backbones, ResNet18 and ResNet34. While we also considered DenseNet121, EfficientNet-B0, and ResNet50, they proved to be too memory-intensive due to the large size of their output features. Even with a reduced batch size of 2, training was unfeasibly slow. Consequently, we settled on utilizing only ResNet34 and ResNet18. Additionally, we performed a center crop of the images to 112 pixels because the cells in most images are centered. Augmentation was also applied through horizontal and vertical flips. The best score was achieved with ResNet34, which obtained a balanced accuracy of **88.16% locally** and **85.714% on the public Kaggle leaderboard**.
As part of an **ablation study**, we attempted to train this model with the original image size (224x224) to assess its impact on model performance. However, this did not improve performance, and training was very slow. For this training, we used a learning rate of 1e-3 and optimized with Adam, utilizing BCEWithLogits as the loss function.

What we tried and failed with the MIL Model? We attempted to train on a fixed number of images per patient with numbers like 10, 50, and 80, but the model struggled to learn and could not surpass 75% of balanced accuracy. We also attempted to weight the BCEWithLogits loss by giving a higher weight to the minority class and a lower weight to the majority class, and vice versa. However, with this approach, the model (ResNet34 backbone) achieved 88.59% of balanced accuracy locally, but on the leaderboard, it dropped to 74.80% and 68.05%, indicating clear overfitting.

Aggregation of results from the best models. Using the decision rules for aggregating the predictions of the best attribute-based model and the best image-based model, the minimum decision rule enables us to increase the balanced accuracy to **86.493% on the leaderboard**.

Decision rule	Min	Max
Leaderboard balanced accuracy	**86.493**	84.935

Table 6: Balanced accuracy sur le leaderboard pour les différentes règles de décision d'agrégation des prédictions des modèles.

5 Discussion and conclusion

In this study, we tackled the task of lymphocytosis classification using various models and methodologies, emphasizing both attribute-based and image-based approaches. Our exploration encompassed traditional machine learning techniques such as logistic regression and decision trees, as well as deep learning architectures like convolutional neural networks (CNNs) and Multiple Instance Learning (MIL) models.

Our findings reveal several insights. Firstly, attribute-based models, particularly neural network-based approaches, demonstrated robust performance, achieving a balanced accuracy of 90% locally and 85.714% on the leaderboard. This underscores the importance of patient attributes such as lymphocyte count, age, and gender in lymphocytosis classification. On the other hand, image-based models presented more challenges. While CNNs and pre-trained models like VGG16 and ResNet were employed for feature extraction, their performance fell short of expectations, with one-to-one and many-to-one models failing to surpass attribute-based models in terms of balanced accuracy. Notably, our exploration of MIL models yielded promising results. Leveraging ResNet architectures for feature extraction, our MIL approach attained a balanced accuracy of 88.16% locally and 85.714% on the public Kaggle leaderboard. Despite encountering challenges such as overfitting and slow training with larger image sizes, the MIL model showcased the potential of incorporating instance-level predictions in medical image analysis. Regarding the aggregation of results from the best attribute-based and image-based models, employing the minimum decision rule proved effective in enhancing balanced accuracy to 86.493% on the leaderboard. This underscores the importance of thoughtful decision rule selection in model ensemble strategies.

In conclusion, our study underscores the value of integrating diverse methodologies in medical image analysis tasks. While attribute-based models showcase strong performance, image-based approaches, particularly those leveraging MIL, offer promising avenues for further exploration and refinement. Future research could focus on fine-tuning model architectures, exploring alternative aggregation strategies, and expanding the dataset to enhance generalization and robustness. Overall, our findings contribute to advancing the field of medical image analysis and hold implications for improving lymphocytosis classification and diagnosis.

References

[1] DIETTERICH, T. G., LATHROP, R., AND LOZANO-PEREZ, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence 89* (1997), 31–71.

[2] DUMONT, M., MARÉE, R., WEHENKEL, L., AND GEURTS, P. Fast multi-class image annotation with random subwindows and multiple output randomized trees. In *International Conference on Computer Vision Theory and Applications* (2009).

[3] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).

[4] ILSE, M., TOMCZAK, J. M., AND WELLING, M. Attention-based deep multiple instance learning, 2018.

[5] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[6] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation 1*, 4 (1989), 541–551.

[8] LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K., AND DOLLÁR, P. Focal loss for dense object detection, 2018.

[9] MARON, O., AND LOZANO-PÉREZ, T. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, M. Jordan, M. Kearns, and S. Solla, Eds., vol. 10. MIT Press, 1998.

[10] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] TOLLES, J., AND MEURER, W. J. Logistic regression relating patient characteristics to outcomes. *JAMA 316*, 5 (2016), 533–534.

[12] VAPNIK, V. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[13] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).