Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis

Mihir Sahasrabudhe, Pierre Sujobert, Evangelia I Zacharaki, Eugénie Maurin, Béatrice Grange, Laurent Jallades, Nikos Paragios, Maria Vakalopoulou

Deep Multi-Instance Learning Using Multi-Modal Data for Diagnosis of Lymphocytosis

Mihir Sahasrabudhe[1], Pierre Sujobert[2,3], Evangelia I. Zacharaki[4], Eugénie Maurin[2], Béatrice Grange[2], Laurent Jallades[2], Nikos Paragios[5], and Maria Vakalopoulou[1]

[1] Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France
✉{mihir.sahasrabudhe,maria.vakalopoulou}@centralesupelec.fr
[2] Université de Lyon, Faculté de Médecine Lyon Sud, Lyon, France
✉{pierre,sujobert,eugenie.maurin,beatrice.grange,laurent.jallades}@chu-lyon.fr
[3] Cancer Research Center of Lyon, INSERM U1052 UMR CNRS 5286, Equipe labellisée Ligue Contre le Cancer, Lyon, France and Hospices Civils de Lyon, Hôpital Lyon Sud, Service d'hématologie biologique, Pierre-Benite, France
[4] University of Patras, Greece
✉ezachar@upatras.gr
[5] Therapanacea, Paris, France
✉n.paragios@therapanacea.eu

Abstract. We investigate the use of recent advances in deep learning and propose an end-to-end trainable multi-instance convolutional neural network within a mixture-of-experts formulation that combines information from two types of data—images and clinical attributes—for the diagnosis of lymphocytosis. The convolutional network learns to extract meaningful features from images of blood cells using an embedding level approach and aggregates them. Moreover, the 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. Our results demonstrate that even the convolutional network by itself is able to discover meaningful associations between the images and the diagnosis, indicating the presence of important unexploited information in the images. The mixture-of-experts formulation is shown to be more robust while maintaining performance via. a repeatability study to assess the effect of variability in data acquisition on the predictions. The proposed methods are compared with different methods from literature based both on conventional handcrafted features and machine learning, and on recent deep learning models based on attention mechanisms. Our method reports a balanced accuracy of 85.41% and outperfroms the handcrafted feature-based and attention-based approaches as well that of biologists which scored 79.44%, 82.89% and 77.07% respectively. These results give insights on the potentials of the applicability of the proposed method in clinical practice. Our code and datasets can be found at https://www.github.com/msahasrabudhe/lymphoMIL.

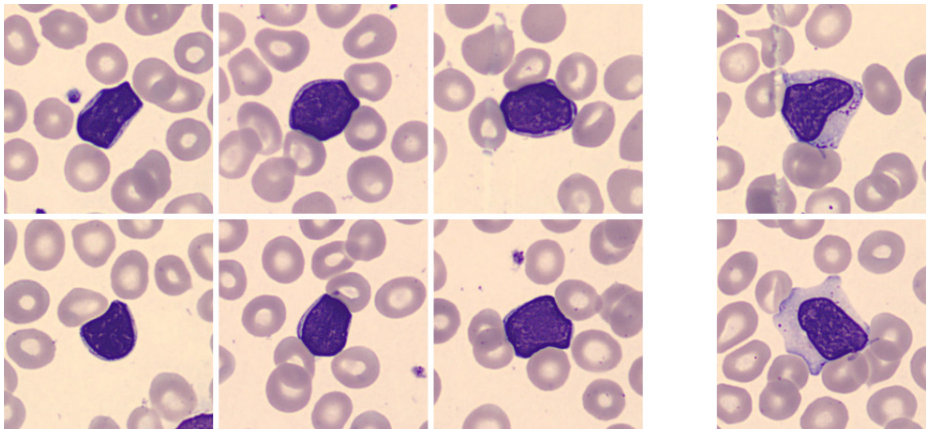


Fig. 1: An example from blood cell images for a patient with a lymphocyte count of 6.967×10^{10}/L. Each image depicts lymphocytes (with dark purple nuclei) surrounded by red blood cells. The six images in the left group depict abnormal lymphocytes while the two in the right one are normal.

1 Introduction

Lymphocytosis (i.e., absolute lymphocyte count above 4×10^9/L) is a common finding, which can be either a reaction to infection, acute stress, and so on (termed *reactive*), or the manifestation of a lymphoproliferative disorder—a type of cancer of the lymphocytes (termed *tumoral*). In existing clinical practice, diagnosis (as either reactive or tumoral) relies on visual microscopic examination of the blood cells (Figure 1) together with the integration of clinical attributes such as age and lymphocyte count. Taking into consideration the visual assessment based on clinical attributes together with texture and size of the lymphocytes in the blood smear, a diagnosis of the subtype of lymphoid malignancy is performed. On the positive side such practice is fast and affordable. It suffers however from poor reproducibility. Additional clinical tests are required, with flow cytometry being the gold standard to definitively affirm the malignant nature of the lymphocytes. However, this analysis is relatively expensive and time consuming, and therefore cannot be performed for every patient in practice. Therefore, the development of automatic and accurate processes could lead to a better way to determine which patient should be referred for flow cytometry analysis, augmenting and assisting the assessment of the clinicians.

Imaging offers great potential to analyze blood cells in a non-invasive and repeatable manner. In the last decade radiomics has emerged in oncology as a way to extract imaging features for diagnosis or prediction of treatment outcome or to be used as a surrogate of oncogenic processes that are difficult to explore by contextual biopsies [1–3]. In a standard radiomic approach, tumors or regions of interest (ROI) are detected and outlined. Features describing, e.g., shape, texture, or morphology, are subsequently extracted [4]. A detailed review of texture analysis methods focusing on microscopy images of cells or tissues can

be found in [5]. In such a setting, (i) the segmented ROIs and pre-defined features are choice-dependent, i.e., only a certain region of the image is used for feature extraction; and (ii) feature extraction is performed independently of statistical modelling, i.e., it is independent of the target training label, thus diminishing the ability to find evidence-driven correlations, a thriving innovation in precision medicine. We argue that we have no evidence that these imaging features capture all correlations between the images and the targets. Moreover, we believe that these independent processes of feature extraction and prediction modelling do not necessarily take full benefit of the richness of all information offered by the images. For our problem setting, modelling these correlations with handcrafted image features becomes even more difficult because of lack of target variables for individual images. The diagnosis being performed over a set of images of varying cardinality, the target variables (reactive and tumoral) are only available for the entire set, i.e., at the patient-level. It is challenging for biologists to annotate each individual image as either normal or abnormal, and yet this annotation suffers from inter-observer variability. Finally, presence of individual abnormal lymphocytes does not guarantee tumoral nature of the symptoms, again making the patient level assessment a necessity.

The images used for this problem (Figure 1) are acquired from blood smears, which are made by placing a drop of blood between two slides in order to create a thin, uniform layer of blood so that individual blood cells are non-overlapping and can be observed under a microscope. These images are then captured using a DM-96 device (Cellavision) while focussing on individual lymphocytes.

In this paper, we present a novel approach for the challenging task of diagnosis of lymphocytosis. Our proposed mehod is able to predict the nature of symptoms (reactive/tumoral) from an acquired set of images of lymphocytes combined optimally with clinical attributes. In particular, the contributions of this paper are fourfold. First, we propose a multi-instance deep convolutional neural network for extracting visual representations from multiple microscopy images and associate them directly with the patient's diagnosis. Second, we investigate how different aggregation methods for the multiple instance scores affect the model's predictions compared to directly trained attention mechanisms. Thirdly, we introduce a mixture-of-experts model [6] in order to learn a classifier from both images and the patient's clinical attributes to render a reliable diagnosis. Finally, we show comparisons with classical image-based methods coupled with multi-instance classification, as well as recent deep learning-based attention methods reporting better performance.

The paper is organized as follows. Section 1.1 discusses previous work on multiple-instance learning as well as deep learning applied to medical analysis. We then describe our method, and present its components and implementation details in Section 2, followed by descriptions of competing methods in Section 3. The dataset used for this study is introduced in Section 4, which is followed by a discussion of the evaluation setting and the results of our experiments. This section also discusses interpretability of the model through saliency visualized using guided backpropagation (Section 5). An extensive comparison with other

methods and ratings of clinical experts is presented along with a discussion of the results in Section 6.

1.1 Related Work

Multiple instance learning (MIL) [7,8] applies to problems where objects (bags) are described by multiple observations (instances) with labels being provided only for the bags. It can be also considered as a form of weakly supervised learning. The challenge that arises for such representations is the lack of precise annotation for each individual instance, and the fact that some of the instances could lack information or encode even misleading information about the object's class (e.g. not all cells are abnormal in a sample of blood smears from a patient exhibiting tumoral behaviour, as shown in Figure 1). While the literature on using machine learning to exploit information contained in cytometry images is sparse, recent work in the analysis of cells and tissue exists, for example, histopathological images. We discuss some such recent advances in this section.

Several methods have been proposed exploiting local or global information and implementing different classifiers or mapping functions [9, 10]. Specifically for histopathological image analysis, a variety of machine learning techniques have been investigated and are exhaustively presented in various reviews [11,12]. Methods based on content-based image retrieval were very commonly used to address this problem [13,14]. Moreover, especially for the task of cell segmentation, a variety of methods have been proposed using bag-of-words [15], support vector machines [16], neural networks [17] or Gaussian mixture models [18]. However, all these methods use predefined hand-crafted imaging features and fail to take full benefit of the domain specificity. A recent work [19] exploits features from tensor decomposition for histopathological diagnosis to address this problem.

There are a lot of studies that adapt CNN models for MIL by using different pooling layers such as the maximum, mean, generalized mean, log-sum-exponentiation (LSE) [20] or the noisy-and function [21]. In particular, [21] presents a CNN architecture which classifies and segments microscopy images using an end-to-end multiple instance scheme. Further, [22] proposes an attention-based multi-instance architecture to classify histophathological images. Several works have explored multiple-instance based frameworks for computer vision tasks, notably for weakly-supervised semantic segmentation. [23] propose using a CNN to predict pixel-wise heatmaps for object classes coupled with an aggregation function for class scores to give an image-level label. The image-level classifier can be trained with negative log-likelihood. [24] localise objects using a CNN pre-trained on ImageNet on top of sliding windows on images to generate training examples for an object detector for Pascal VOC object detection problem. In a follow-up work [25], they use a modified loss function which transfers labels to sliding windows based on the image-level label. Several studies propose using multi instance learning methods for histopathological image analysis [26–29]. Some of them also investigate the fusion of histopathology images with genomic or molecular data. Both studies presented in [30, 31] propose the concatenation of deep learning based features extracted from the convolutional

layers with the additional data before the fully connected layers of the network. A similar fusion strategy is used in the recent work [32] which aggregates several networks trained on different modalities. Our approach extends the notions of these methods towards a generic multi-instance deep learning framework from weak annotations that are augmented by information relevant to the patient's clinical data. We propose an images-only model as well as a mixture-of-experts model for fusion of these two modalities.

Few works on automated diagnosis of lymphocytosis using machine learning can be found in literature. Two recent works explore the use of clinical data for direct prediction of a lymphoproliferative disorder (also termed *tumoral*). The authors of [33] train a decision tree based on examination results of patients, which they treat as feature vectors. In a more recent work [34], the authors test several classifier models, for example, support vector machines, multi-layer perceptrons, decision trees, and nearest neighbour algorithms, on cell population data and clinical attributes, and predict three classes—healthy control, viral infection, and chronic lymphocytic leukemia. However, both of these works do not directly use images to train their classifiers. To the best of our knowledge, ours is the first work to train a deep learning model using images coupled with clinical data for the diagnosis of lymphocytosis.

2 Deep multi-instance learning

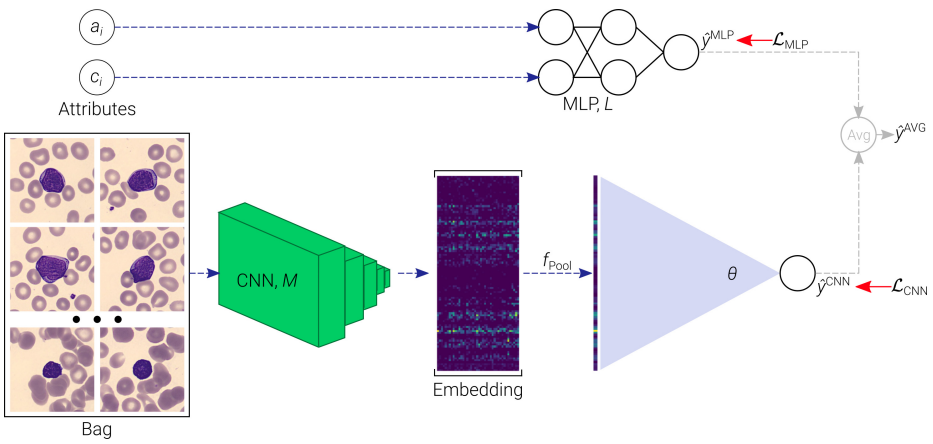
In this section, we present our deep learning-based framework for the task of predicting the presence of a lymphoproliferative disorder. First, we briefly discuss all the notations used and the main MIL scheme. This is followed by a description of all the components and details of the proposed framework. We follow this with a brief review of competing methods. Experimental results and comparisons with these methods follow. We conclude with a discussion of the findings.

2.1 Notation

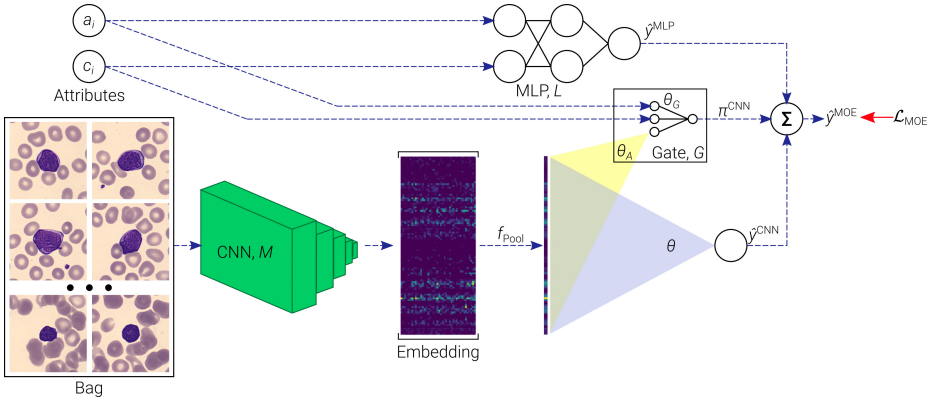
Let us first introduce some notation to describe the proposed approach. We are given a set of N subjects, with a set of images being associated with each patient—the number of which can vary from one patient to the other—along with patient attributes, namely age and lymphocyte count. We represent the data of a patient as

$$S_i = \left(\left\{ X_i^j \right\}_{j=1}^{j=N_i}, a_i, c_i, y_i \right), \tag{1}$$

where $\{X_i^j\}$ represents the N_i images obtained from the i-th patient, and a_i, and c_i represent their age in years, and lymphocyte count in number of cells per litre of blood, respectively. The target class is represented by a binary variable $y_i \in \{0, 1\}$, with the values indicating a reactive and tumoral nature, respectively. We will use this notation throughout the paper, while further pertinent notation shall be introduced later.



(a) CNN, MLP, and AVG models



(b) Mixture-of-Experts (MOE) model

Fig. 2: Schematic representations of the models—*top*: the CNN, MLP, and AVG models; and *bottom*: the MOE model. Please see text for a detailed explanation. Red arrows in the AVG model indicate that data flow through these arrows is not involved in the training phase, but only in the prediction phase. Further, $\mathcal{L}_{CNN}$ and $\mathcal{L}_{MLP}$ in the top figure indicate where these training losses are applied for the CNN and MLP models. In the bottom figure, Σ refers to Equation 17.

2.2 Standard MIL assumption

In the standard MIL assumption each instance is considered to fall into one of the two categories—positive (1), or negative (0) [8, 35]. Furthermore, the existence of one or more positive class instances in the bag renders the bag itself positive. Concretely, this can be written as

$$y_i = \begin{cases} 1 & \text{if } \sum_j y_i^j \geq 1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since $y_i^j \in \{0, 1\}$, this equation can further be simplified as

$$y_i = \max_j y_i^j \,. \tag{3}$$

This, however, is quite a strong assumption for our problem. Firstly it requires knowledge of the instance-level class which is not available. To address this challenge, many algorithms arbitrarily assume that each instance inherits the class from the bag it belongs to. Such an assumption is not suitable for our problem because the malignancy of an individual lymphocyte is uncertain looking only at the blood smear as cytologists can have differing opinions on the matter. Secondly, the presence of only one *abnormal* lymphocyte does not justify a diagnosis of tumoral behaviour. To adhere more closely to the knowledge-based diagnostic approach followed by cytologists, we should draw inference from all instances in a bag.

This reasoning leads us to the choice of a more general aggregation approach. The aggregation function should be one that is invariant to permutation of the instances. Broadly, we can classify aggregation approaches into two classes instance-level and embedding-level MIL [35]-

1. **Instance-level MIL**. This approach aggregates instance-level predictions to give bag-level predictions. Thus, a model predicts y_i^j, which is followed by an aggregation function to yield an estimate of y_i. Examples of aggregation functions that fall into this category are the max and mean functions, log-sum-exp [36], log-mean-exp, noisy-or [37], and noisy-and [21].
2. **Embedding-level MIL**. In this approach, instead of aggregating predictions at the instance-level, a low-dimensional embedding of instances is learnt, and a bag-level classifier is trained on top of the aggregation of the embeddings of all instances in the bag. We shall refer to the vector resulting after the aggregation as the *pooled feature vector*, and the aggregation operation itself as *pooling*. This approach was employed in [38], and also shown to perform well on document classification [39, 40], as well as whole-slide histopathology images for discriminative patch detection [41], and nuclei localization [22]. This approach more closely models our problem, wherein the aggregation function serves the purpose of accumulating and summarizing knowledge obtained from all available blood smears.

It can easily be observed that the max approach discussed above is indeed invariant to permutation, and is an instance-level approach. In this paper, we employ a deep-learning model with embedding-level pooling. The premise for using an embedding-level approach is based on the earlier discussion indicating that the standard multi-instance learning assumption is not suitable for the diagnosis of lymphocytosis.

2.3 Proposed Deep Learning Architecture

The proposed deep learning architecture consists of a convolutional neural network as a feature extractor. The CNN works on the entire (unmasked) lymphocyte images. The final aim of the proposed framework is the accurate prediction

of the probability $P(y_i = 1 \mid S_i)$, where the variable $y_i = 1$ indicates the presence of disease, and $y_i = 0$, otherwise. In this study, we introduce a deep learning model which draws patient-level inference using only the images $\{X_i^j\}$ to model this probability. For such a setting, each patient will be referred to as a *bag*, and the y_i-s as *bag-level labels*. Similarly, each image in $\{X_i^j\}$ will be referred to as *instance* and the y_i^j will denote the *instance-level* labels corresponding to X_i^j. For the training of the model, the only provided annotations are y_i.

CNN for Blood Smears A convolutional feature extractor is used to generate low-dimensional embeddings for each of the instances. The feature extractor is followed by a pooling operation in the embedding space, and a classifier which predicts the probability of disease trained on top of the pooled representations. We design the model so that it is end-to-end trainable, in that it learns the low-dimensional embedding as well as the classifier jointly (Figure 2a).

Let M be the function that represents this feature extractor. M operates on instances (X_i^j) and generates an embedding in a low-dimensional space. Let f_{Pool} represent a pooling function on these embeddings which is permutation-invariant. In this study, we investigate three pooling functions—the element-wise maximum function (f_{Max}), the element-wise average function (f_{Mean}), and the log-sum-exp function (f_{LSE}). These are defined as

$$f_{\text{Max}}\left(\{\mathbf{h}_i^j\}\right) = \left(\max_j \mathbf{h}_i^j(k)\right)_{1 \leq k \leq E}; \tag{4}$$

$$f_{\text{Mean}}\left(\{\mathbf{h}_i^j\}\right) = \frac{1}{N_i}\sum_j \mathbf{h}_i^j \text{ ; and} \tag{5}$$

$$f_{\text{LSE}}\left(\{\mathbf{h}_i^j\}\right) = \frac{1}{r}\log\left(\frac{1}{N_i}\sum_j \exp\left(r \cdot \mathbf{h}_i^j\right)\right), \tag{6}$$

where $\mathbf{h}_i^j = M(X_i^j) \in \mathbb{R}^E$ represents the embedding of X_i^j, and E is the dimension of this embedding. The pooled embeddings over all instances are further represented by the vector $\mathbf{p}_i$, i.e.,

$$\mathbf{p}_i = f_{\text{Pool}}\left(\{\mathbf{h}_i^j\}\right), \tag{7}$$

for $f_{\text{Pool}} \in \{f_{\text{Max}}, f_{\text{Mean}}, f_{\text{LSE}}\}$. We use a ResNet [42] as our feature extractor M. We, however, set the width of the ResNet as a hyperparameter of our model. Denoting by K, the the base "step size", shown in Table 1 is the architecture of the ResNet, with the number of channels doubling at each residual layer. In a standard ResNet [42], K is set to 64. However, the original ResNets were intended for large-scale computer vision applications, and as such use "wide" latent representations. Since our problem has limited data, we make a design choice to experiment with different values of K, where $K \in \{8, 16, 32, 64\}$.

Once the aggregated representation of a bag is generated as an embedding in the low-dimensional space, a linear classifier is used to predict the bag label. The linear classifier assigns a score to the bag given by

$$\hat{y}_i^{\text{CNN}} = \boldsymbol{\theta}_C^\top \mathbf{p}_i + \beta \,, \text{and} \tag{8}$$

$$P(y_i = 1 \mid \{X_i^j\}) = \sigma\left(\hat{y}_i^{\text{CNN}}\right) . \tag{9}$$

Here, $\sigma(x) = \dfrac{1}{1+\exp(-x)}$ is the logistic function, and $\boldsymbol{\theta}_C$ and β are, respectively, the weight vector and the bias of the classifier. Overall, this models the bag probability as a Bernoulli distribution. Negative log-likelihood is used to train the model end-to-end. Concretely, the loss function is defined as

$$\mathcal{L}_{\text{CNN}} = -\ln \sigma(\hat{y}_i^{\text{CNN}}) \,\big|_{y_i=1} - \ln\left(1 - \sigma(\hat{y}_i^{\text{CNN}})\right) \big|_{y_i=0} \,. \tag{10}$$

We will henceforth refer to this model as CNN (in small caps). It should be noted that CNN uses only the images for diagnosis.

Multi-layer Perceptron for Clinical Data Since the clinical data (a_i, c_i) are also helpful to the cytologists during diagnosis, we employ them in our model as an additional source of information. In order to integrate the clinical data, a multi-layer perceptron is used consisting of one hidden layer and one output layer to predict the probability of disease.

The multi-layer perceptron consists of an input layer with two units, connected to a hidden layer which also has two units. The sigmoid activation function is used in the hidden layer. The output layer has just one unit which represents the score of the classifier (Figure 2a). The score for a bag i is denoted by $\hat{y}_i^{\text{MLP}}$. This multi-layer perceptron is also trained with the negative log-likelihood loss as described in Equation 10. Let L represent the multi-layer perceptron. Then we can write

$$\hat{y}_i^{\text{MLP}} = L(a_i, c_i) \,; \tag{11}$$

$$P(y_i = 1 \mid a_i, c_i) = \sigma(\hat{y}_i^{\text{MLP}}) \,; \text{ and} \tag{12}$$

$$\mathcal{L}_{\text{MLP}} = -\ln \sigma(\hat{y}_i^{\text{MLP}}) \,\big|_{y_i=1} - \ln\left(1 - \sigma(\hat{y}_i^{\text{MLP}})\right) \big|_{y_i=0} \,. \tag{13}$$

We will henceforth refer to this model as MLP. It should be noted that MLP does not use images for diagnosis.

We have now described two models that use different training data to predict the same variable. The two types of input data are not completely independent of each other. Based on this, the predictions of the two models are combined in two possible ways.

Averaging Model The averaging model simply averages the two scores from these two models. The combined prediction is

$$\hat{y}_i^{\text{AVG}} = \frac{1}{2}\left(\hat{y}_i^{\text{CNN}} + \hat{y}_i^{\text{MLP}}\right) . \tag{14}$$

Since each predictor can be trained separately, there is no joint training in the averaging model. We refer to this model as AVG.

Mixture-of-Experts Model So far, the probability of tumoral nature was modelled using either the blood smears (CNN model), or the clinical attributes (MLP model), with no sharing of these informations between the models. It is not unreasonable to assume that the two models might have disagreements over certain examples, as the biologists themselves are not always in agreement. It therefore makes sense to *choose* the better of the two models depending on the patient. To this end, a mixture-of-experts [43–45] model is studied to learn simultaneously from both, the images as well as the clinical attributes. However, instead of targeting cooperation between the models, in which the loss function over the average of both models' predictions is minimized, we wish to promote specialisation, such that each model specializes over a certain set of examples [6]. More concretely, two "experts"—the CNN and the MLP— are employed together with a gating network weighting the contributions of the two experts (Figure 2b). The gating network operates on the pooled features $\mathbf{p}_i$, as well the attributes a_i and c_i, and outputs a set of mixing coeffecients. Such a model learns to output a mixture of probability distributions learnt by each of the experts. Examples of uses of a mixture-of-experts are applications to speech recognition [6, 46, 47] and disease classification [48], among other tasks.

The gating network is formulated as an aggregation kernel learned on the embedding space, followed by a linear layer to regress the contributions. The complete model used for the gating network is

$$\pi_i^{\text{CNN}} = G(\mathbf{p}_i, a_i, c_i) = \sigma\left(\boldsymbol{\theta}_G^\top \begin{bmatrix} \boldsymbol{\theta}_A^\top \mathbf{p}_i \\ a_i \\ c_i \end{bmatrix}\right) \text{ ; and} \tag{15}$$

$$\pi_i^{\text{MLP}} = 1 - \pi_i^{\text{CNN}}\,, \tag{16}$$

where π_i^{CNN} and π_i^{MLP} contributions of the CNN and the MLP, respectively. The final prediction of the mixture-of-experts model is given by

$$P(y_i = 1 \mid S_i) = \hat{y}_i^{\text{MOE}} = \pi_i^{\text{CNN}} \sigma\left(\hat{y}_i^{\text{CNN}}\right) + \pi_i^{\text{MLP}} \sigma\left(\hat{y}_i^{\text{MLP}}\right)\,, \tag{17}$$

The mixture-of-experts model uses the gating network parameterized by $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_G$, and hence can be trained end-to-end with the the two experts. The loss function employed to train this model encourages specialisation, in that it lets each expert concentrate on examples it can classify better. Concretely, the loss function is formulated as

$$\mathcal{L}_{\text{MOE}} = -\ln \hat{y}_i^{\text{MOE}} \Big|_{y_i=1} - \ln\left(1 - \hat{y}_i^{\text{MOE}}\right) \Big|_{y_i=0}\,. \tag{18}$$

We will henceforth refer to the mixture-of-experts model as MOE. Three different paradigms are further explored in the MOE framework—(P1) training the entire model end-to-end with no initialisation; (P2) initialising the CNN and MLP with models trained uniquely with $\mathcal{L}_{\text{CNN}}$ and $\mathcal{L}_{\text{MLP}}$, respectively, and then training only G; and (P3) initialising the CNN and MLP as before, and training end-to-end.

Layer	Layer Name	Output Size	Residual Blocks
1	conv1	112×112	-
2	conv2	56×56	$\begin{bmatrix} 3 \times 3, K \\ 3 \times 3, K \end{bmatrix}$
3	conv3	28×28	$\begin{bmatrix} 3 \times 3, 2K \\ 3 \times 3, 2K \end{bmatrix}$
4	conv4	14×14	$\begin{bmatrix} 3 \times 3, 4K \\ 3 \times 3, 4K \end{bmatrix}$
5	conv5	7×7	$\begin{bmatrix} 3 \times 3, 8K \\ 3 \times 3, 8K \end{bmatrix}$
6	flatten	$7 \cdot 7 \cdot 8K$	-

Table 1: Architecture of the convolutional neural network M used to estimate $\mathbf{h}_i^j$. The input to the network is an image of size 224×224. Each row defines an operation, where each convolution is followed by batch normalisation and rectified linear unit (ReLU). The residual layers are layers 2-5. K is a hyperparameter which determines the width of the residual network. We test with the values $\{8, 16, 32, 64\}$ for K.

2.4 Training

Our networks are trained with the negative log likelihood loss. Depending on the used models, we employ one of the losses out of $\mathcal{L}_{\text{CNN}}$, $\mathcal{L}_{\text{MLP}}$, and $\mathcal{L}_{\text{MOE}}$ to evaluate each of the components used in this study. Training is performed with standard backpropagation. We observed that large batch sizes result in a much more stable model than using a batch size of 1 (as in [22]). Several models were trained with different combinations of configurations, i.e., with varied combinations of K, f_{Pool}, training data (images, attributes), and averaging and mixture-of-experts models. While training the CNN, we do not randomly draw a set of images of a fixed size, but instead use all images corresponding to a patient.

Overfitting As there are very few training examples, we find that the model is susceptible to overfitting. To reduce overfitting, standard data augmentation is introduced during training. Random horizontal and vertical flips are added along the x- and y-axes, as well as random rotations from the set $\{0°, 90°, 180°, 270°\}$.

Further, standard colour augmentation is employed [49]. In particular, principal component analysis (PCA) is performed on RGB pixel values over the training dataset. Then for a training image, three values, α_i, are sampled from a normal distribution with mean 0 and standard deviation 0.1. The colour of the training image is then rescaled by adding

$$[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3] \left[\alpha_1 \lambda_1, \alpha_2 \lambda_2, \alpha_3 \lambda_3\right]^\top \tag{19}$$

where $\mathbf{e}_i$ and λ_i are, respectively, the eigenvectors and eigenvalues of the 3×3 covariance matrix of RGB pixel values over the entire training dataset.

Finally, we record the performance of the model on the validation set at each training epoch (Figure 3).

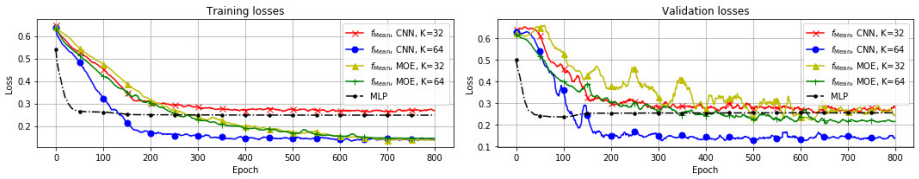


Fig. 3: Training and validation losses for the f_{Mean} pooling function for varying ResNet widths.

2.5 Implementation Details

The code was written in Python with the PyTorch library [50], and executed on a machine equipped with a NVIDIA GTX 1080 GPU, a 12-core 3.5 GHz processor, and 32 gigabytes of memory. The models were trained using the Adam optimiser [51], starting with a learning rate of 0.0001, and decay it by a factor of 0.1 every $96,000$ iterations. We use $\beta_1 = 0.9$ and a weight decay of 0.0005. Training is done for $220,000$ iterations, in which the learning rate is decreased by a factor of 0.1 twice. While the number of training iterations is fixed, we also choose the best model found during training according to its performance on the validation set. Training one model takes about one and a half days.

The original images in our dataset are of size 360 pixels $\times$ 360 pixels, but we resize them to 224 pixels $\times$ 224 pixels, as we observed that we do not lose any significant information under the resizing operation, and it allows us to curb overfitting as well as use less memory overall. The RGB values of images are then centered using the per-pixel mean over the entire dataset.

3 Compared Methods

3.1 Learning from clinical attributes

We first compare our method against standard classifiers applied on the clinical attributes. In particular, we train an SVM, decision trees, and a Gaussian naive Bayes classifier, as well as ensemble models like AdaBoost, gradient boosting classifier, and random forests. We show the results of these experiments in Tables 3 and 4.

3.2 Classical Approach

To evaluate the performance of our proposed method we compare it with an MIL framework using classical imaging features.

Feature Extraction Before we employ an MIL scheme, a feature extraction step on the blood smear images is required. These features must be extracted from the area of interest, i.e., the lymphocytes in the images for our case. Under this framework, a segmentation of lymphocytes in needed to compute several imaging and shape characteristics. The lymphocytes were automatically segmented in each image instance X_i^j using the Simple Linear Iterative Clustering (SLIC) superpixels algorithm [52], after a smoothing operation as a preprocessing step. SLIC is a gradient ascent method implementing a local K-means clustering to generate a K-superpixel segmentation. Since the best value for K is not known in advance, we perform multiple segmentations for different values of K and then created an average segmentation for each instance and each of the RGB image channels. This multi-scale fuzzy segmentation step did not require any parameter tuning and aimed to smooth the boundaries of ambiguous regions while at the same time retain the crisp boundaries of regions that were present in more scales. The smoothed RGB image was then segmented using the K-means clustering algorithm in HSV (hue, saturation, value) colour representation scale using $K = 3$. The three obtained clusters represented *i)* the lymphocytes (with the cytoplasm), *ii)* all other cells, and *iii)* the background. If several lymphocytes were found in an image, only the largest of them was retained and used for feature extraction. The analysis of the classical image characteristics was based on 94 features extracted from each of the segmented blood smear images per subject. These features are described in detail below.

Shape The shape of the largest lymphocyte in each image was described by 12 features: area, major axis length, minor axis length, eccentricity, convex area, filled area, Euler number, equivalent diameter, solidity, extent, and perimeter calculated in two ways using different weights for diagonal pixels and corners.

Image Statistics 3 intensity statistics (minimum, maximum, average) were extracted for each of the 3 RGB channels inside the region of interest.

Texture 24 texture variables [53] including the average fractal dimension and statistical measures (autocorrelation, contrast, correlation$_1$, correlation$_2$, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity$_1$, homogeneity$_2$, maximum probability, sum of squares, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation$_1$, information measure of correlation$_2$, inverse difference, normalised inverse difference, moment normalised inverse difference) from the gray-level co-occurrence matrix were calculated for pairs of pixel in 0°, 45°, 90°, 135°, for each of the 3 channels.

Density The number of lymphocytes in the image was used as a measure of cell density.

MIL Training The feature vectors from all blood smear images of each subject comprised a multiple instance dataset which was introduced into a MIL

classifier for prediction of lymphocytosis. We investigated several standard MIL classifiers from the multiple instance learning literature, such as the expectation maximization maximum diverse density (EMDD) [54], multi-instance support vector machine (MI-SVM) [55], multi-instance learning in embedded subspaces (MILES) [56], but the best performing, which was finally selected, was the specialising MIL (SPEC_MIL) which is a generalisation of MI-SVM. The only hyperparameter in this algorithm is the fraction of positive instances, which was tuned by 3-fold cross validation on the training set and then fixed to the value attaining most often the highest classification accuracy. Two experiments were performed. The one relied only on the handcrafted image features whereas the other included also a_i and c_i. Integration of the clinical variables with the imaging features was performed in an early phase and led to a joint dataset that was introduced into the multi-instance classifiers.

Model	$K=8$			$K=16$			$K=32$			$K=64$		
	f_{Max}	f_{Mean}	f_{LSE}	f_{Max}	f_{Mean}	f_{LSE}	f_{Max}	f_{Mean}	f_{LSE}	f_{Max}	f_{Mean}	f_{LSE}
CNN	0.60	0.82	0.82	0.47	0.88	0.72	0.47	0.95	0.77	0.89	**0.96**	0.82
AVG	0.87	0.90	0.91	0.88	0.91	0.89	0.88	0.92	0.89	0.90	0.94	0.90
MOE-P1	-	0.90	-	-	0.92	-	-	0.94	-	-	0.94	-

Table 2: Area under the receiver operating characteristics curve for various models and configurations. Similarly, CNN is trained only using the images. MOE-P1 refers to from-scratch training of the MOE model (Section 2.3).

3.3 Attention-based Methods

We also experimented with the attention-based model recently proposed in [22]. In this approach, a CNN is trained along with an attention mechanism which learns to focus on discriminative images in data. This approach employs a pooling function in the latent space which is effectively a weighted average, with the weights being determined by softmax attention. We invite the reader to refer to [22] for further details on the model.

4 Dataset

To build a dataset for this problem, blood smears and patient attributes were collected from 204 patients from the routine hematology laboratory of the Lyon Sud University Hospital. The samples were anonymized as required by the General Data Protection Regulation, keeping basic demographic information (age and sex) intact. The inclusion criteria were (a) a lymphocyte count above 4×10^9/L, and (b) absence of opposition to the research. The blood smears were automatically produced by a Sysmex automat tool, and the nucleated cells were automatically photographed with a DM-96 device (Cellavision). All the cells labelled

Method	Data	Sensitivity	Specificity	Accuracy	Balanced Accuracy
Biologists	imgs, attrs	0.7529 ± 0.0953	0.7885 ± 0.0126	0.7639 ± 0.0690	0.7707 ± 0.0705
AdaBoost	attrs	0.7586	0.6923	0.7381	0.7255
SVM	attrs	0.8621	0.6923	0.8095	0.7772
Decision Trees	attrs	0.7241	0.6154	0.6905	0.6698
Random Forest	attrs	0.8276	0.6154	0.7619	0.7215
Gaussian Naive Bayes	attrs	0.5517	0.8462	0.6429	0.6989
Gradient Boosting	attrs	0.6897	0.6923	0.6905	0.6910
MLP	attrs	0.8621	0.6923	0.8095	0.7772
Classical Approach	imgs	**1.0000**	0.3846	0.8095	0.6923
	imgs, attrs	0.8966	0.6923	0.8333	0.7944
DeepMIL [22], $K = 32$	imgs	0.9655	0.6923	**0.8810**	0.8289
DeepMIL [22], $K = 64$	imgs	**1.0000**	0.2308	0.7619	0.6154
f_{Mean}, $K = 32$, CNN	imgs	0.9310	0.6923	0.8571	0.8117
f_{Mean}, $K = 32$, AVG	imgs, attrs	0.9310	0.6923	0.8571	0.8117
f_{Mean}, $K = 32$, MOE-P1	imgs, attrs	0.8621	**0.8462**	0.8571	**0.8541**
f_{Mean}, $K = 32$, MOE-P2	imgs, attrs	0.8621	0.6923	0.8095	0.7772
f_{Mean}, $K = 32$, MOE-P3	imgs, attrs	0.8621	0.6154	0.7857	0.7387
f_{Mean}, $K = 64$, CNN	imgs	0.8621	**0.8462**	0.8571	**0.8541**
f_{Mean}, $K = 64$, AVG	imgs, attrs	0.9310	0.6154	0.8333	0.7732
f_{Mean}, $K = 64$, MOE-P1	imgs, attrs	0.8621	**0.8462**	0.8571	**0.8541**
f_{Mean}, $K = 64$, MOE-P2	imgs, attrs	0.8621	**0.8462**	0.8571	**0.8541**
f_{Mean}, $K = 64$, MOE-P3	imgs, attrs	0.8966	0.6923	0.8333	0.7944

Table 3: Different evaluation metrics for models discussed in this paper, evaluated on the testing cohort for the diagnosis of lymphocytosis. The second column signifies the type of incorporated training data (imgs: images, attrs: clinical attributes, i.e. a_i and c_i), as explained in section 2.4.

as lymphocytes by the DM-96 device were used for analysis. To determine the presence of a disorder, flow cytometry was used incorporating a panel of antibodies for the diagnosis of lymphoproliferative disorders (CD3, CD4, CD5, CD8, CD10, CD56, CD20, CD19, kappa, lambda). The results of this test were used as the ground truth for the presence or absence of tumoral behaviour. In our dataset, the minimum and maximum number of images per patient were 16 and 198, with a mean and standard deviation of 82 and 45, respectively.

The 204 patients were divided into training, validation and test sets. The training cohort used of all our models consists of 142 subjects with 44 reactive and 98 malignant cases. The validation cohort consists of 21 subjects with 6 reactive and 15 malignant cases, while the test cohort includes 42 subjects with 13 reactive and 29 malignant examples.

5 Experimental Results

In this section, we compare the results of the proposed with the competing methods, as well as with the visual assessment annotations from 12 different biologists from the Lyon University Hospital. The biologists provided their diagnoses for each of the patients of the test cohort, based on the images and the supporting clinical data. The obtained results are evaluated and compared based on the following metrics: sensitivity, specificity, accuracy, balanced accuracy and in terms of area under receiver operating characteristic curve (ROC-AUC).

In Table 2 different components of the proposed method are evaluated in terms of ROC-AUC. The tested aggregation functions and the width of ResNet

(K) are evaluated for the CNN, AVG, and MOE-P1 models. The best performance is obtained by $K = 64$ and using the f_{Mean} pooling operation for both models. Based on these observations we performed the evaluation of the MOE model only for the f_{Mean} operation.

In Table 3 a comparison between the different methods is presented, including an ablation study for the different components (MLP, CNN and MOE) of our study. To calculate the evaluation metrics we did not perform any optimisation for the threshold value and a value of 0.5 is used to separate reactive and tumoral cases for all the methods. In general, the predictions of the biologists have a very wide variation that can reach even 9% for the sensitivity metric. Moreover, we can observe that a lot of information is captured by patient attributes as standard classifiers are at par with the average performance of the experts in almost all of the evaluation metrics and reach similar balanced accuracy. This is also indicated in classical image characteristics as our experiments indicate a boost in the overall and balanced accuracy when the attributes are combined with the predefined features extracted from the images. However, the performance of the classical approach is inferior to the one reported by the attention based method [22]. The latter obtains quite high sensitivity but relatively low specificity indicating that this method detects much more false positives for the diseased category.

Table 3 summarises also the performance of our proposed method using different configurations and parameters. Most of them outperform all baselines with all the metrics, while all of them are higher than 70%. In particular, different configurations of the proposed method namely the $f_{\text{Mean}}, K = 32$, MOE-P1, $f_{\text{Mean}}, K = 64$, CNN and $f_{\text{Mean}}, K = 64$, MOE-P1 report the highest balanced accuracy while they also report very high values for the rest of the metrics. This demonstrates the robustness of the method on different configurations. It also noteworthy that the proposed model based solely on imaging information can perform similarly with models that use additional source of information about the patients.

For a better and more complete evaluation of the reported methods we compare the areas under their ROC curves (ROC-AUC) in Table 4. In general all the methods report ROC-AUC greater than 0.83 with the proposed method using the f_{Mean} reporting values higher than 0.91 proving its robustness and stability. Finally, the models that are based only on clinical information report 0.89 as their highest ROC-AUC which is at least 3% lower than the models that use information produced by the images.

5.1 Repeatability

In order for our system to be used in clinical practice, it needs to be robust in terms of repeatability. That is to say, the proposed models should arrive at the same conclusion as long as a clinically relevant set of images is sampled from a patient for testing. To this end, we design a test to assess the performance of our model over several image sets sampled from the same blood sample of a patient. From five additional patients participating in the study, five new images

Model	ROC-AUC
Biologists' average	0.9204
AdaBoost	0.7255
SVM	0.7771
Decision Trees	0.6698
Random Forest	0.7215
Gaussian Naive Bayes	0.6989
Gradient Boosting	0.6910
MLP	0.8912
Classical Approach, imgs	0.8727
Classical Approach, imgs+attrs	0.8329
Deep MIL [22], $K = 32$	0.9151
Deep MIL [22], $K = 64$	0.9390
$f_{\text{Mean}}, K = 32$, CNN	0.9416
$f_{\text{Mean}}, K = 64$, CNN	**0.9629**
$f_{\text{Mean}}, K = 32$, MOE-P1	0.9416
$f_{\text{Mean}}, K = 32$, MOE-P2	0.9178
$f_{\text{Mean}}, K = 32$, MOE-P3	0.9151
$f_{\text{Mean}}, K = 64$, MOE-P1	0.9443
$f_{\text{Mean}}, K = 64$, MOE-P2	0.9178
$f_{\text{Mean}}, K = 64$, MOE-P3	0.9390

Table 4: A comparison of the CNN, MOE models using the f_{Mean} pooling function with attention models by ROC-AUC together with the comparisons with the attention module, classical approach and the MLP model trained only with the patient attributes. We also show a comparison against the average prediction of the twelve biologists.

sets were extracted. As the images are extracted from the same blood sample, the diagnosis using each image set should be the same. The goal of this test of repeatability is then to evaluate the performance of the proposed models per smear for each patient and examine the variance that is introduced in them.

In Table 5, we list the result of the best CNN, MLP, and MOE models. The true behaviour for each patient is listed in row 2, while the prediction of each of the two models is listed in rows 4, 5, and 6. For the *prediction* row of CNN and MOE-P1, each column indicates which image set (row 3) was used for diagnosis, whereas the *maj. vote* row is the diagnosis obtained by a majority vote over the predictions on the image sets. We note that the MOE model is more stable in terms of its conclusion with much fewer intra-patient disagreements, whereas the images-only CNN model is more sensitive to the set of sampled images as there are more intra-patient disagreements. By majority vote, the MOE model is also able to give the correct prediction for each patient while the CNN and MLP models fail.

Patient		A	B	C	D	E
Ground truth		Reactive	Reactive	Reactive	Tumoral	Tumoral
Image set		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
cnn	prediction	R T R T T	T T T T T	T R R R T	T T T T T	T T T T T
	maj. vote	Tumoral	Tumoral	Reactive	Tumoral	Tumoral
mlp	prediction	Reactive	Tumoral	Tumoral	Tumoral	Reactive
moe-P1	prediction	R R R R R	R T R T R	R R R R R	T T T T T	T R T T T
	maj. vote	**Reactive**	**Reactive**	**Reactive**	**Tumoral**	**Tumoral**

Table 5: Test of repeatability. Five different sets of images were obtained from each of five patients. The patient number, the ground truth, and the corresponding image sets are shown in the first three rows, while the corresponding predictions are in the last three rows. CNN *prediction* and MOE *prediction* indicate the result of these models for the corresponding image set in row 3. *maj. vote* is the diagnosis obtained from majority voting over these five predictions. Since the MLP prediction is independent of the image set used, there is only one result per patient using this model. T refers to a diagnosis of tumoral behaviour, while R refers to one of reactive.

5.2 Interpretability

We asked an expert from the *Hospices Civils de Lyon* to highlight regions in lymphocyte cells that are important for manual diagnosis. In Figure 4, we show these two regions. In particular, the inside of the cytoplasm and the nucleus, as well as the shape of their borders influence the decision of whether a cell is tumoral or not. We then applied guided backpropagation [57] coupled with mul-

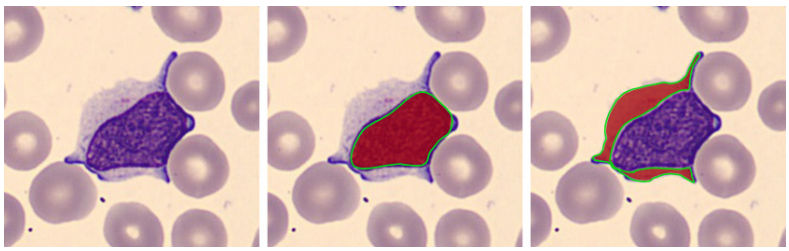


Fig. 4: Important regions in a lymphocyte for diagnosis. *left*: an example lymphocyte image; *centre*: inside the nucleus and its shape/border; and *right* inside the cytoplasm and its shape/border.

tiplication with the input [58] on our model to visualize areas of the cells termed

important by the network for diagnosis. In some cases, the most informative pixels correspond to sub-cellular structures which are also used by cytologists to decipher the nature of the lymphocytes. For example, the appearance of the nucleolus is strongly indicative of a tumoral lymphocyte (Figure 5 (a) and (b)), whereas the presence of small granulations in the cytoplasm is characteristic of cytotoxic T or NK lymphocytes observed in reactive situations (Figure 5 (c) and (d)).

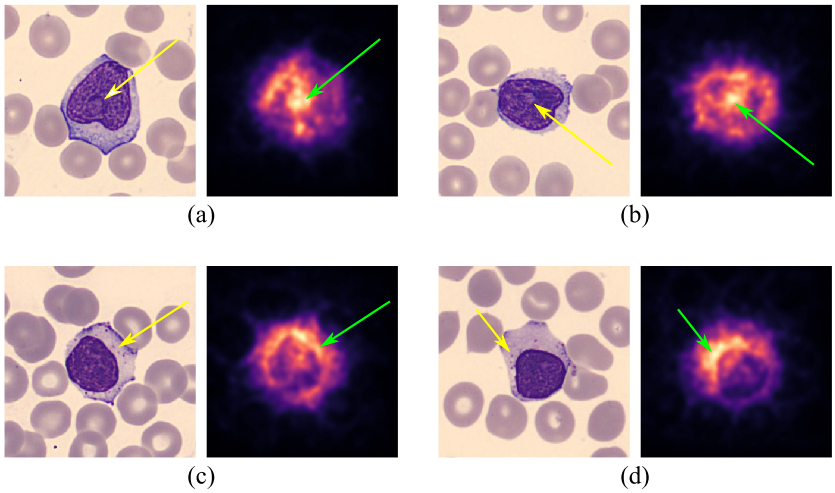


Fig. 5: Saliency using guided backprop times input [58]. Yellow arrows overlaid on lymphocytes represent discriminative features in the image, while green arrows represent the corresponding locations in the saliency maps. In (a) and (b), the arrows point at nucleoli in the cells, while in (c) and (d), they show granulations in the cytoplasm.

6 Discussion

To the best of our knowledge, this is the first study that provides a deep learning-based method for accurate diagnosis of lymphocytosis. Our proposed method is compared with standard multi instance learning schemes that are used in literature and other recently proposed deep learning based methods while it is also compared with the visual assessment of 12 different biologists. Our experiments indicate the superiority of our method, showing the potential of such a tool for automated diagnosis of lymptocytosis in clinical practice.

Different pooling operators, network parameters (number of channels K), configurations (using images and clinical attributes) and training strategies are reported in this study in order to show the behavior of our proposed method.

Starting with the pooling operators, f_{Max} and f_{LSE} show comparatively poor performance over the tested values of K. We postulate that the f_{Max} operator is too strong for the problem at hand, while the f_{LSE} operator—being a smooth approximation to the maximum—performs better than f_{Max}. However, both f_{Max} and f_{LSE} fail to capture the relationship between the instance and the bags unlike f_{Mean}, which reports the best performance for all the configurations.

Our experiments further indicate that the models work better when a wider ResNet is used, i.e., a higher value of K. In particular, narrower networks seem not powerful enough to capture all the available information and learn enough features from the data. This is in accordance with other studies in literature [59]. However, both $K = 32, 64$ perform similarly with both values reporting very close performance (Tables 2 and 3). This result also alleviates overfitting concerns with wider models.

Concerning the training strategies used for the MOE model, our experiments indicate that the MOE model works better with the P1 training paradigm, where both models are trained end-to-end without initialisation. This is expected behaviour under the training loss used. The training is done to encourage specialisation, but under the P2 and P3 paradigms, the participating experts (CNN and MLP) have been pre-trained to fit the entire data instead of specialising over a portion of it, whereas under the P1 paradigm, they are uninitialised.

We argue that the attention-based models [22], which are one of the competing methods, have lower performance than the proposed approach because of the high variance in the number of images per patient. As the attention-based models use a softmax function to compute image weights, these weights tend to become skewed when there are several *important* examples in the set. This renders the learning the classifier a more difficult task.

Here, it is worth mentioning that almost all the methods reach similar and higher performances compared to the experts indicating the high potentials of such a tool in clinical practice. However, it should be also noted that the biologists evaluated digital images of the blood smears, and not directly the blood smears under a microscope. This might have lowered their performances, because most of them work usually with a microscope. Our experiments also show that a deep learning-based method is able to extract more discriminative features than a classical approach. The performance of the images-only model (CNN) shows that it is possible to extract and exploit information from blood smears using an automated tool. However, it is still sensitive to the set of images extracted as seen in Section 5.1. The MOE model, on the other hand, is able to correct errors that the CNN and MLP models were making individually. This indicates that while neither of clinical attributes and images alone is enough to make a reliable diagnosis, the MOE model is able to combine information from them for the correct diagnosis. This demonstrates the robustness of the MOE model to data acquisition. While the repeatability of the model is satisfying but not perfect, as the whole preanalytical workflow is automated, it is technically and economically feasible to increase the number of blood smears analyzed per patient in order to increase accuracy.

Finally, another aspect that can be taken into account is the time-efficiency of the proposed approaches. The proposed methods are fast when drawing inference, making the entire process rapid and efficient. For our test cohort which contains 42 patients, testing required 30s in all, which corresponds to about 0.72s per test example. This time is better than the one needed by a biologist who may need considerably longer for the examination of one case.

7 Conclusion and Future Work

This study presents a deep MIL scheme for reliable diagnosis of lymphocytosis. Imaging features from lymphocytes which are extracted automatically are coupled with patient attributes in a dynamic way taking advantage of all available information for each patient. Our method has been validated under different training schemes and different pooling operators proving its robustness and accuracy. Moreover, it has been also evaluated against human experts, classical handcrafted features combined with MIL frameworks and recently proposed attention-based methods. We also propose a mixture-of-experts model which combines information from acquired blood smears and clinical attributes of a patient for a more robust assessment. Overall, we found that deep learning based approaches outperform conventional methodologies while models that are based only on the images report better performance demonstrating their diagnostic capacity for lymphocytosis prediction. A repeatability test also evaluates the robustness of the CNN and the MOE models and demonstrates that the MOE model is indeed able to combine information from the two sources efficiently (attributes and images) for a more reliable diagnosis.

As we can also see from Table 3, our method outperforms the biologists' average prediction. With all our experiments, we demonstrated that our method can give a reliable tool to biologists in order to assist them in their everyday practice, being deployed in real-life scenarios. However, further tests, especially using datasets from different hospitals, must be undertaken in order to extensively validate the accuracy of the method. This constitutes the second and clinically significant part of our future work.

Acknowledgements. We thank Stergios Christodoulidis for his valuable comments.

Funding. This research had been partially supported by the Partner University Fund and the the Fondation pour la Recherche Medicale (FRM; no. DIC20161236437).

References

1. D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, Jan 2017.

2. E. J. Limkin, R. Sun, L. Dercle, E. I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, E. Deutsch, and C. Ferté, "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology," *Annals of Oncology*, vol. 28, no. 6, 2017.
3. R. Sun, E. J. Limkin, M. Vakalopoulou, L. Dercle, S. Champiat, S. R. Han, L. Verlingue, D. Brandao, A. Lancia, S. Ammari, A. Hollebecque, J.-Y. Scoazec, A. Marabelle, C. Massard, J.-C. Soria, C. Robert, N. Paragios, E. Deutsch, and C. Ferté, "A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study," *The Lancet Oncology*, vol. 19, no. 9, 2018.
4. E. Zacharaki, S. Wang, S. Chawla, D. Yoo, R. Wolf, E. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme," *Magnetic Resonance in Medicine*, vol. 62, no. 6, 12 2009.
5. S. D. Cataldo and E. Ficarra, "Mining textural knowledge in biological images: Applications, methods and trends," *Computational and Structural Biotechnology Journal*, vol. 15, 2017.
6. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, Mar. 1991.
7. J. D. Keeler, D. E. Rumelhart, and W. K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in *Advances in neural information processing systems*, 1991, pp. 557–563.
8. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Prez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, 1997.
9. J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81 – 105, 2013.
10. J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, 03 2010.
11. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, 2009.
12. D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, 2018.
13. X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards large-scale histopathological image analysis: Hashing-based image retrieval," *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, 2015.
14. R. Sparks and A. Madabhushi, "Out-of-sample extrapolation utilizing semi-supervised manifold learning (ose-ssl): Content based image retrieval for histopathology images," *Scientific Reports*, vol. 6, pp. 1455–1462, 06 2016.
15. J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 2009.
16. F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, 2016.
17. N. Theera-Umpon and S. Dhompongsa, "Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, 2007.

18. M. Dundar, S. S. Badve, G. Bilgin, V. C. Raykar, R. K. Jain, O. Sertel, and M. N. Gürcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 1977–1984, 2011.
19. T. Papastergiou, E. Zacharaki, and V. Megalooikonomou, "Tensor decomposition for multiple-instance classification of high-order medical data," *Complexity*, 12 2018.
20. J. Ramon and L. D. Raedt, "Multi instance neural networks," *ICML workshop on attribute-value and relational learning*, 2000.
21. O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.
22. M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 2127–2136.
23. P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.
24. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.
25. ——, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 685–694.
26. X. Zhu, J. Yao, F. Zhu, and J. Huang, "Wsisa: Making survival prediction from whole slide histopathological images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7234–7242.
27. E. Wulczyn, D. F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament-Auvigne, C. H. Mermel, P.-H. C. Chen, Y. Liu, and M. C. Stumpe, "Deep learning-based survival prediction for multiple cancer types using histopathology images," *PLoS One*, vol. 15, no. 6, p. e0233678, 2020.
28. Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan *et al.*, "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4837–4846.
29. Y. Xu, J.-Y. Zhu, I. Eric, C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical image analysis*, vol. 18, no. 3, pp. 591–604, 2014.
30. P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. V. Vega, D. J. Brat, and L. A. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018.
31. J. Yao, X. Zhu, F. Zhu, and J. Huang, "Deep correlational learning for survival prediction from multi-modality data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 406–414.
32. R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, 2020.

33. L. Wang, Y. Wang, Y. Chen, C. Liu, and X. Fan, "Prediction of lymphocytosis using machine learning algorithm based on checkup data," in *2017 4th International Conference on Systems and Informatics (ICSAI)*. IEEE, 2017, pp. 649–654.
34. L. Bigorra, I. Larriba, and R. Gutiérrez-Gallego, "Machine learning algorithms for accurate differential diagnosis of lymphocytosis based on cell population data," *British journal of haematology*, vol. 184, no. 6, pp. 1035–1037, 2019.
35. J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, 2010.
36. J. Ramon and L. De Raedt, "Multi instance neural networks," in *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, 2000, pp. 53–60.
37. O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.
38. X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.
39. M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, and N. de Freitas, "Modelling, visualising and summarising documents with a single convolutional neural network," University of Oxford, Tech. Rep. arXiv:1406.3830, 2014. [Online]. Available: http://arxiv.org/abs/1406.3830
40. D. Kotzias, M. Denil, P. Blunsom, and N. de Freitas, "Deep multi-instance transfer learning," *arXiv preprint arXiv:1411.3128*, 2014.
41. L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.
42. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
43. J. B. Hampshire and A. Waibel, "The meta-pi network: building distributed knowledge representations for robust multisource pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 751–769, July 1992.
44. M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.
45. P. Moerland, "Mixtures of experts estimate a posteriori probabilities," in *International Conference on Artificial Neural Networks*. Springer, 1997, pp. 499–504.
46. S. J. Nowlan and G. E. Hinton, "Evaluation of adaptive mixtures of competing experts," in *Advances in neural information processing systems*, 1991, pp. 774–780.
47. S. R. Waterhouse, "Classification and regression using mixtures of experts," Ph.D. dissertation, Citeseer, 1998.
48. S.-K. Ng and G. J. McLachlan, "Extension of mixture-of-experts networks for binary classification of hierarchical data," *Artificial Intelligence in Medicine*, vol. 41, no. 1, pp. 57–67, 2007.
49. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
50. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.
51. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

52. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, 2012.
53. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.
54. Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *Advances in neural information processing systems*, 2002, pp. 1073–1080.
55. S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.
56. Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.
57. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
58. A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.
59. S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, September 2016, pp. 87.1–87.12.