Lymphocytosis classification
Team Name: Iberdeep

Biel Castaño-Segade BIEL.CASTANO_SEGADE@ENS-PARIS-SACLAY.FR
David Faget-Caño DAVID.FAGET_CANO@ENS-PARIS-SACLAY.FR

Introduction

The goal of this project is to identify lymphoproliferative disorders —a category of cancer— in individuals with lymphocytosis, marked by an elevated count of lymphocytes.

The primary challenge lies in discerning this condition, as lymphocytosis frequently occurs in patients and may simply signal a benign response to various factors, such as infections. Differentiating between benign and malignant causes of lymphocytosis remains a significant challenge in clinical practice. Conventional methods, such as visual inspection of blood smears, together with taking into account the patient's clinical data, often lack the necessary precision. Other methods like flow cytometry offer increased accuracy, but their cost and limited availability hinder widespread adoption. This project seeks to address these limitations by developing a new automated system to assist clinicians in identifying patients who might require further evaluation through flow cytometry.

To achieve this goal, we used a dataset where the anonymized blood smears and the corresponding demographic information were collected from 204 patients diagnosed with elevated lymphocyte counts at Lyon Sud University Hospital. The dataset is segmented, with 142 patients allocated for training the model and the remaining 42 reserved for testing.

Our proposed model takes a multifaceted approach, incorporating both image data from blood smears and patient metadata. A pre-trained ResNet34 (He et al., 2015) convolutional neural network (CNN) will be employed to extract key features from the blood smear images. To summarize the information contained in the sequence of images associated with a patient, an attention mechanism will be used to aggregate the features of the images. Furthermore, a separate branch within the model will process patient metadata. By combining these elements and passing them through a classifier, we aim to develop a robust and informative model.

In this project, we will not only develop the model architecture but also delve deeper to optimize its performance. We will explore various strategies for hyperparameter selection, validation methodology, preprocessing of images, and model configurations.

1. Architecture and Methodological Components

1.1. Architecture

1.1.1. Motivation

In the context of our problem, leveraging both detailed image features and patient metadata (such as age, gender, and lymphocytes count) is important for accurate predictions. This

task requires handling multiple images (not always the same number) per patient and combining these with non-image data to make a single prediction per patient. When treating a patient's sequence of images as a bag, this approach falls under multiple instance learning, a method previously utilized in medical imaging research. (Hou et al., 2015; Kraus et al., 2016).

1.1.2. Model Components and Their Rationale

In this section, we delve into the various components of our model. In summary, our network architecture integrates two distinct branches: one dedicated to extracting features from the sequence of images, and another focused on extracting features from the clinical data. Following the feature extraction, these two sets of features are concatenated and used by a final classifier to generate predictions. All these components are trained at the same time. To elaborate further:

i. **Pre-trained Image Feature Extractor:** To extract features from each image in a patient's sequence, we employ deep convolutional neural networks pre-trained on ImageNet. By using these pre-trained models through transfer learning, especially in our case with limited data available, we can significantly enhance performance —a strategy widely adopted in numerous Kaggle challenges-. In particular, we chose a pre-trained ResNet34 (He et al., 2015), which offers a more complex architecture than ResNet18 (as we will present later) and showed a better ability to capture important features from blood smear images effectively. This approach lets us avoid training our model from scratch, minimizing the risk of overfitting and improving the model's generalization capabilities.

ii. **Attention Module as image feature aggregation:** After extracting features from each image in a patient's sequence, we need to aggregate the features to create a single set of features. We chose an attention mechanism as the aggregation method as was done in (Ilse et al., 2018). The motivation for this arises from the observation that not all images, or even regions within an image, contribute equally to a final diagnosis. This mechanism allows the model to prioritize the most informative features across multiple images, mirroring the selective focus approach of a radiologist who naturally emphasizes certain features over others. The underlying intuition is that by dynamically assigning weights to different images based on their content, the attention module empowers the model to aggregate features effectively.

   It is important to note that this is an attention mechanism in the sense that it allows our model to focus more on certain inputs than on others based on learned weights. However, it's a specific use case of attention, designed for aggregating features from multiple instances, and it does so in a simpler and more specific manner compared to the transformers attention mechanisms (Dosovitskiy et al., 2020).

   For implementing it, we apply a sequence of operations, a linear transformation followed by a Tanh activation to project the input features to a hidden dimension, another linear transformation to reduce it to a single weight per image, and a Softmax function to normalize these weights across all images in a batch. The module

then uses these weights to perform element-wise multiplication with the original input, aggregating the results to highlight the most relevant features across multiple images.

iii. **Patient Metadata Processing** Our motivation for introducing a separate processing branch for patient metadata, such as age, stems from our data exploration findings which revealed its significant influence on diagnoses. This metadata, being structurally distinct from image data, necessitates dedicated handling. The intuition behind having a dedicated neural network branch for metadata is to enable the model to learn and extract relevant representations from non-image data.

The implementation begins with a linear transformation from 3 to 64 dimensions, followed by a ReLU activation to introduce non-linearity. This pattern of a linear layer followed by ReLU activation is repeated twice more, with each sequence maintaining the 64-dimensional space. Dropout layers with a probability of 0.2 are introduced to prevent overfitting by randomly omitting a fraction of the units during training.

iv. **Final Classifier** Since the challenge lies in effectively combining heterogeneous data sources (image features and metadata) for a unified prediction, we must concatenate the aggregated image features with metadata features to provide a good representation of each patient's case. The final classification layer may then use this combined feature set to make a prediction, benefiting from the multimodal context.

By integrating an attention mechanism with a multi-instance learning framework and adding a dedicated branch for metadata processing, the model is tailored to selectively focus on the most pertinent information across both images and metadata.

1.2. Other Architectural experiments

We conducted experiments using ResNet18 and ResNet34 as feature extractors and observed that ResNet34 consistently outperformed ResNet18, improving both validation and test balanced accuracy by approximately 5%. Additionally, experiments comparing mean aggregation to our attention aggregation mechanism revealed that the attention mechanism significantly improved performance, with mean aggregation methods yielding balanced accuracies that were 6% lower. Removing the patient metadata processing component also resulted in a decrease in accuracy, approximately 13% lower, underscoring the value of incorporating patient metadata into our model. These findings will be detailed further in the ablation study section. We also tried another more sophisticated aggregation mechanism, based in attention layers like the ones used in the transformer architecture, but we were not able to make it learn well.

2. Model tuning and comparison

2.1. Preprocessing

Inspired by (Sampathila et al., 2022), that used a very similar dataset, we noticed an increased saturation of lymphocytes within the images. This observation led us to apply a mask by thresholding the saturation channel in the HSV color space, resulting in images where pixels not corresponding to lymphocytes were rendered black (Fig. 1). This technique

was adopted to enhance the discernibility of lymphocytes, potentially elevating our model's ability to accurately identify relevant features. However, contrary to our expectations, Fig. 2 shows that the model actually performed approximately 10% better on average (again in terms of validation and test balanced accuracy) without this preprocessing step. It appears that the additional focus on saturation and specific image regions did not align well with the model's learning objectives, highlighting the importance of aligning preprocessing techniques with the specific characteristics and requirements of the model and task at hand. Therefore, we have chosen to keep only a resize of the image to 224x224.

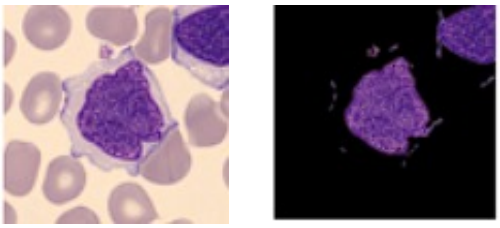

Figure 1: Visual comparison of images without (left) and with (right) preprocessing.

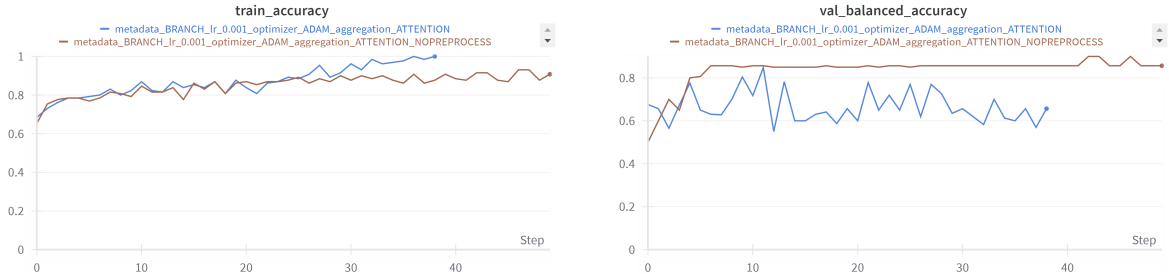

Figure 2: Metrics with and without preprocessing. Training accuracy is better with preprocessing, validation balanced accuracy is not. This is because the model overfits the training data.

2.2. Training and hyperparameters choosing procedure

During the training phase of our selected model and architecture, we used the framework Weights & Biases (wandb) for detailed logging of training and validation metrics. Our focus was on hyperparameter optimization, where we tested various learning rates (0.001, 0.005, 0.01) and optimizers (SGD, ADAM), aiming for the highest balanced accuracy on the validation dataset. The Binary Cross-Entropy loss function was chosen for its suitability in binary classification tasks, effectively penalizing incorrect predictions. To refine training, we used an exponential learning rate scheduler with gamma set to 0.9, reducing the learning rate progressively to enhance model convergence. An early stopping mechanism capped training at 50 epochs to prevent overfitting. Our final model selections were based on superior balanced accuracy results on both validation and public test sets, with the best

model employing a 0.001 learning rate and the ADAM optimizer. Training was executed on a Kaggle notebook with dual T4 GPUs and each execution took approximately 2 hours.

2.3. Results

2.3.1. Validation procedure and results

We employed a stratified approach to create the validation set, considering both age and label to reflect the strong correlation between age distribution and label occurrence, as revealed by our boxplot analysis. This method ensures a balanced representation of each age range within every label category. Additionally, we split the data into training and validation sets following an 80/20 proportion. This split balanced the need for sufficient training data with the requirement for a robust validation set to evaluate our model's performance accurately. Metric values achieved by our best model can be found in Table 1.

Metric	**Train Acc**	**Train Bal Acc**	**Val Acc**	**Val Sens**	**Val Spec**	**Val Bal Acc**
Values	93%	90%	94%	92%	80%	86%

Table 1: Summary of the best model performance metrics

2.3.2. Test set results

The limited size of both the validation and public test sets led to some fluctuations in balanced accuracy across our model trainings. Nonetheless, our two top-performing executions achieved notably high accuracies of 0.93246 and 0.89870 in the public test set.

2.3.3. Ablation study

We carried out several ablation studies to test the impact of different configurations on our model's performance. Initially, we experimented with bypassing the dedicated neural network for processing metadata and directly inputting the metadata into the final classifier. This had a negative impact on the validation balanced accuracy, diminishing it from 86% to 73%. Next, we explored the effect of aggregating features using a simple mean operation as opposed to employing an attention mechanism, which also diminished the validation balanced accuracy from 86% to 80%.

Lastly, we evaluated our model's performance by separately considering only the image data and only the metadata. In both scenarios, we observed a reduction in the validation balanced accuracy by more than 10% (57% and 64% of validation balanced accuracy, respectively), indicating the combined use of both data types significantly enhances model balanced accuracy.

Conclusion

This document emphasizes the primary methodology we adopted to tackle the data challenge presented in the Deep Learning for Medical Imaging course. By focusing on the iterations, assumptions, dataset representations, and model choices, we identify major obstacles in accurately differentiating between reactive and tumoral lymphocytosis. Our most effective model achieved a score of 0.93 on the public academic leaderboard.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

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

Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. *CoRR*, abs/1504.07947, 2015. URL http://arxiv.org/abs/1504.07947.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018. URL http://arxiv.org/abs/1802.04712.

Oren Z. Kraus, Jimmy Lei Ba, and Brendan J. Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, June 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw252. URL http://dx.doi.org/10.1093/bioinformatics/btw252.

Niranjana Sampathila, Krishnaraj Chadaga, Neelankit Goswami, Rajagopala P. Chadaga, Mayur Pandya, Srikanth Prabhu, Muralidhar G. Bairy, Swathi S. Katta, Devadas Bhat, and Sudhakara P. Upadya. Customized deep learning classifier for detection of acute lymphoblastic leukemia using blood smear images. *Healthcare*, 10(10), 2022. ISSN 2227-9032. doi: 10.3390/healthcare10101812. URL https://www.mdpi.com/2227-9032/10/10/1812.