DLMI - Lymphocytosis classification Kaggle Challenge

Imane SI SALAH IMANE.SI_SALAH@ENS-PARIS-SACLAY.FR

Ali Ahmadi ALI.AHMADI@ENS-PARIS-SACLAY.FR

Abstract

Lymphocytosis, a common hematological condition, can indicate either a reactive response or lymphoproliferative disorders, including malignancies. Current diagnostic methods rely on manual blood smear examination, supplemented by clinical factors like age and lymphocyte count, leading to subjectivity and limited reproducibility. To address this, our project aims to develop an automated process to distinguish between reactive and tumoral lymphocytosis. Using deep learning techniques and a dataset of 204 individuals, we trained models to classify lymphocytosis cases, achieving a balanced accuracy of 0.85714. By automating classification, we aim to improve diagnostic accuracy and streamline patient referral for further analysis, enhancing hematology practice.

Keywords: Lymphocytosis , lymphoproliferative disorder , medical image , Deep learning , vision Transformer.

1. Introduction

Lymphocytosis, characterized by an abnormal increase in lymphocyte count, presents a diagnostic challenge in clinical practice. While it can indicate various conditions from reactive responses to serious underlying disorders, current diagnostic methods rely on manual examination of blood smears, leading to subjectivity and scalability issues. In response, our project aims to develop an automated classification system using machine learning techniques. Leveraging a dataset from Lyon Sud University Hospital, we aim to accurately differentiate between reactive and tumoral lymphocytosis by integrating image analysis with clinical data. This approach seeks to enhance diagnostic accuracy, streamline decision-making, and assist clinicians in identifying patients needing further evaluation or treatment.

2. Methodology and Data

2.1. Data Collection and Preprocessing

2.1.1. Data Source

The dataset originates from the routine hematology laboratory of the Lyon Sud University Hospital. It comprises anonymized blood smears and patient attributes collected from 204 individuals who met some inclusion criteria, such as having a lymphocyte count above 4×10*9/L and providing consent for research purposes.

2.1.2. Dataset Description

The dataset includes basic demographic information such as age and sex, along with clinical attributes like lymphocyte count, which are essential for lymphocytosis diagnosis. Each patient is associated with a folder containing blood smear images captured by a Sysmex automat tool, providing visual information for analysis.

Note that: **'0' denotes the reactive cases and '1' the cancerous ones**

2.1.3. Data Visualization and analysis

The dataset under study presents a variety of features across a total of 163 patients. It encompasses patient identifiers, binary class labels, gender, date of birth, lymphocyte counts, and age.

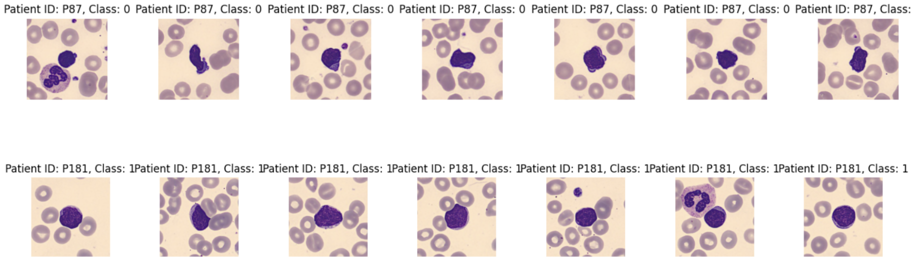
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P87, Class: 0
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1
Patient ID: P181, Class: 1

Figure 1: some sample images form class 0 and class 1 respectively.

- Class Distribution The first visualization (Figure 2) shows the distribution of the binary class label. The majority class '1' includes 113 patients, indicating a prevalence of this class over the minority class '0', which consists of 50 patients.
- Distribution of Age The histogram of age (Figure 3) suggests a quasi-normal distribution with a moderate right skew, indicating an aging cohort predominantly between 60 and 80 years old.
- Age Distribution by Class Label The boxplot of age by class label (Figure 6) reveals that the median age of class '0' is lower than class '1'. The broader interquartile range for class '1' suggests a greater spread of ages within this group.
- Age Distribution by Gender The comparative boxplot of age by gender (Figure 5) shows that the median age is relatively consistent between the genders. However, the female cohort exhibits a slightly wider age range.
- Distribution of Gender Lastly, the gender distribution (Figure 4) shows a nearly equal number of males (82) and females (81).

2.1.4. Data Preprocessing

Class Balancing: As You can see in Figure 11 , there is Imbalance images distribution per patient, we see that images per patient range from 16 images for patient P62 to a max 198

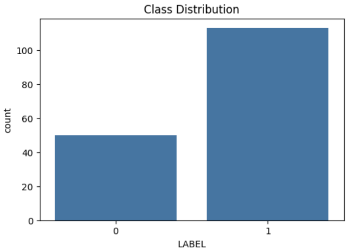

Figure 2: Class distribution of the dataset.

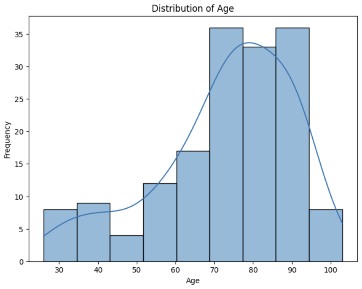

Figure 3: Histogram of the age distribution.

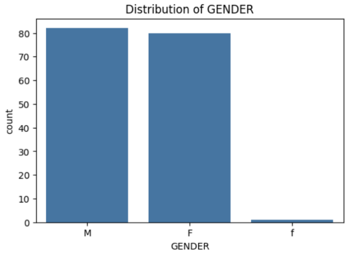

Figure 4: Bar chart of gender distribution.

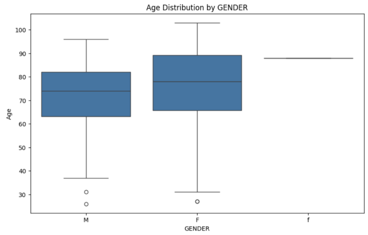

Figure 5: Box plot of age by gender.

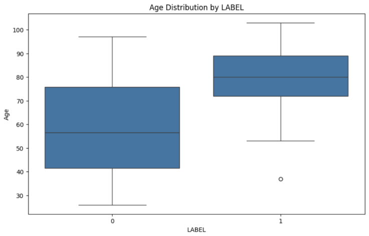

Figure 6: Boxplot of age by class label.

Figure 7: Visualizations depicting the class, age, and gender distributions within the dataset.

images for patient P35, in addition as we can see from the initial analysis of the dataset, there is a significant class imbalance, which is a common issue in medical datasets where one class significantly outnumbers the other. As illustrated in Figure 2, the majority class (Label 1) had a substantially higher count compared to the minority class (Label 0).

To address this we employed two strategies, the first is to up sample the minority class, by repeating some images as shown in Figure 8, then instead of upsampling we attempted down sampling of the majority class instead although this approach was not good since it involves a major loss of information as shown in Figure 9. Overall, we noticed that both approaches end up in decreasing the performance of the model since it overfits the data.

These preprocessing steps help to mitigate the potential bias towards the majority class and improve the robustness of the subsequent machine learning model. By training the model on a dataset with a balanced class distribution, we aim to enhance its ability to generalize and perform accurately on unseen data.

Data Augmentation:

To boost our model's performance on new data, we applied data augmentation techniques like resizing images to 256x256 pixels, randomly flipping them, adjusting brightness

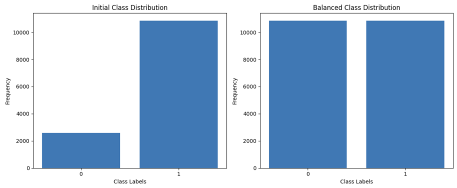

Figure 8: Oversampling the minority class

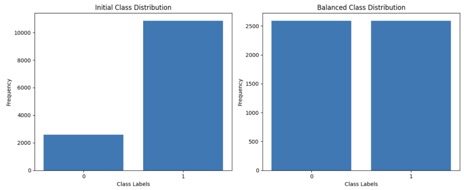

Figure 9: Undersampling the majority class

Figure 10: Balanced class distributions after oversampling and undersampling.

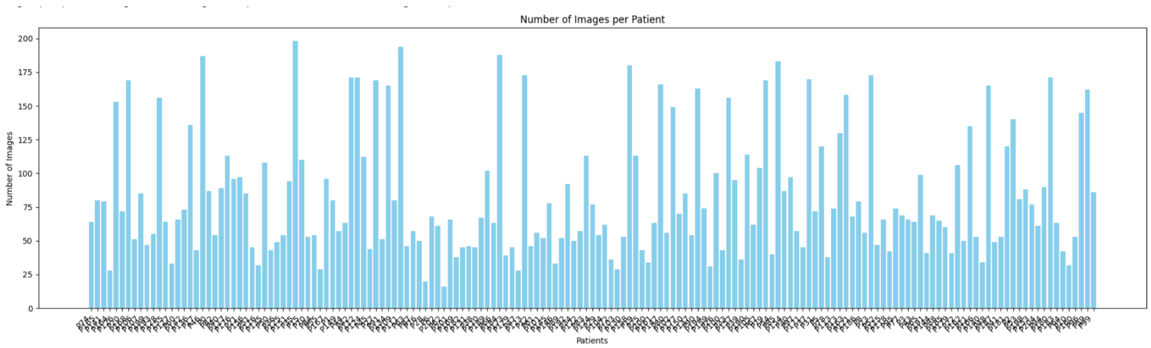

Figure 11: there is a clear imbalance between how the images are distributed per patiens

and contrast, rotating by up to 30 degrees, and applying affine and perspective transformations. These steps helped our model handle a variety of image conditions better.

By augmenting the training dataset with these diverse transformations, we aimed to expose the model to a wide range of variations present in real-world images. This approach not only helps prevent over fitting but also enhances the model's ability to generalize to unseen data, ultimately leading to improved classification performance and clinical utility.

2.2. Evaluation Metrics

This competition is evaluated on balanced accuracy. The balanced accuracy, commonly used in classification problems, normalizes true positive and true negative predictions by the number of positive and negative samples, respectively, and divides their sum by two. In particular, if sensitivity is defined as TPR and specificity as TNR:

$$\text{BalancedAccuracy} = \frac{TPR + TNR}{2}, \text{ where } TPR = \frac{TP}{TP + FN}, TNR = \frac{TN}{TN + FP}$$

2.3. Experiments and results

We used several models in our quest for the best model, we summarize the main architectures we used:

2.3.1. First model

The first model implemented is a Convolutional Neural Network (CNN) architecture designed for binary classification tasks. It consists of three convolutional layers followed by activation and max-pooling operations in order to extract features then we used 2 fully connected layers and finally a sigmoid layer as the classification part. we trained this model using the AdamW optimizer with learning rate of 0.01 and Binary Cross Entropy (BCE) loss. we obtained a validation balanced accuracy score of: **85%**

The loss and accuracy plots are given in Figure 12

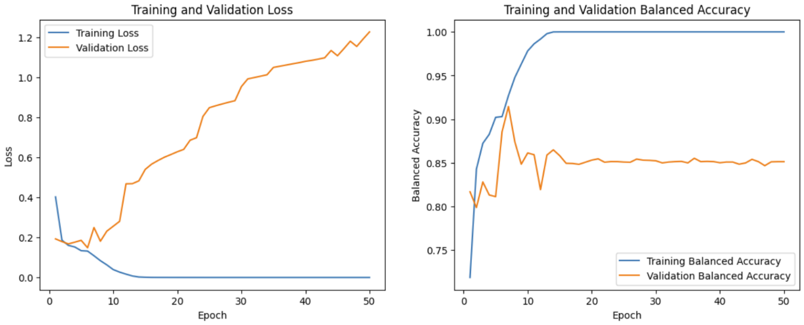

Figure 12: Loss and balanced accuracy plots for model 1

We notice a weird upward trend in the validation loss after few iterations , indicating that the model does not generalize well and clearly overfits the training set.

In addition, we also trained this model using the balanced dataset, for both oversampling and under sampling methods, we obtained a balanced validation score of **92%** and **89%** respectively, loss and accuracy curves are displayed in figure13.

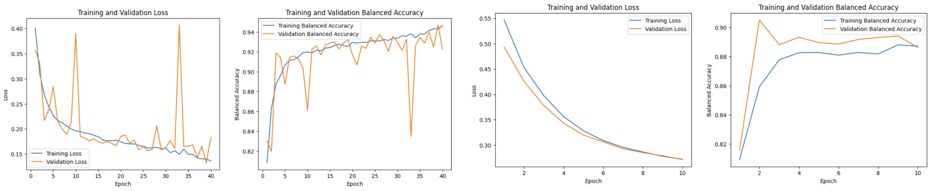

Figure 13: Loss and balanced accuracy plots for model 1.

From the results on the loss and accuracy we see that the accuracy of the model trained on over sampled minority class is very unstable .

2.3.2. Second Model

In an attempt to improve performance, a more complex CNN architecture is introduced. This model expands upon the previous architecture by adding additional convolutional layers and incorporating dropout regularization between layers to prevent overfitting.

In addition, we trained this model with an augmented dataset using augmentation techniques presented in section 2.1.4, for the same optimize and loss as for model 1, we obtained **90%** validation accuracy, and the results are displayed in Figure 14

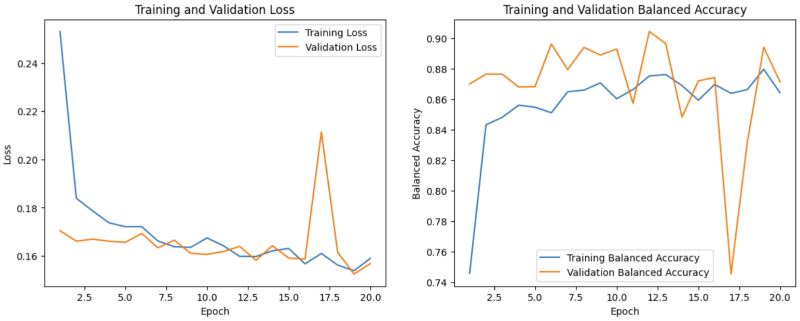

Figure 14: Loss and balanced accuracy plots for model 2

Here, we see a big instability at epoch 17, characterized by a big increase in the validation loss and the a correspoding decrease in the validation score.

2.3.3. Third Model

For this model, we were inspired form the fine tuning techniques seen in the class and the lab, more precisely, we took a pretrained ViT model, that we attempted to further finetune using the AdaptFormer techniques (Chen et al., 2022), that consists on injecting a small bottleneck adapter module, consisting of down-sampling and up-sampling layers and a scaling factor, parallel to the existing feed-forward network in the ViT as shown in figure 15.

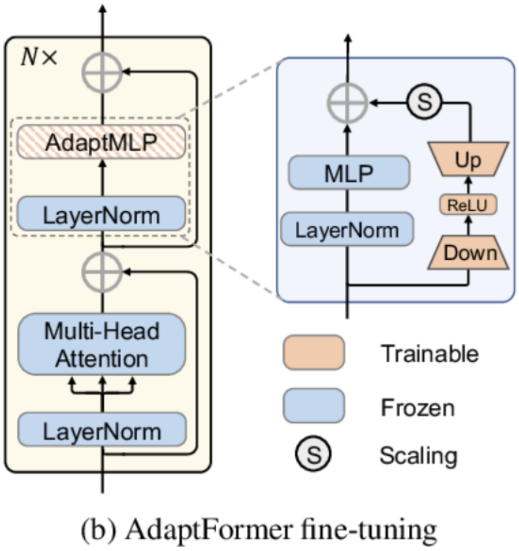

Figure 15: AdaptFormer

Model	Validation score	Test score on Kaggle
Baseline CNN	85.13%	**85.71%**
Baseline CNN with Balanced data	92.29%	83.11%
Enhanced CNN	89.41%	79.74%
Fine Tuned ViT (AdaptFormer)	84.31%	**85.71%**

Table 1: Summary of results

This allows the model to maintain its weights while adapting to the task at hand namely: Lymphocytosis classification.We used the DINOv2 model form (Oquab et al., 2023; Darcet et al., 2023).We were able to obtain a validation accuracy of **85%**, and the Loss and accuracy plots are displayed in figure 16

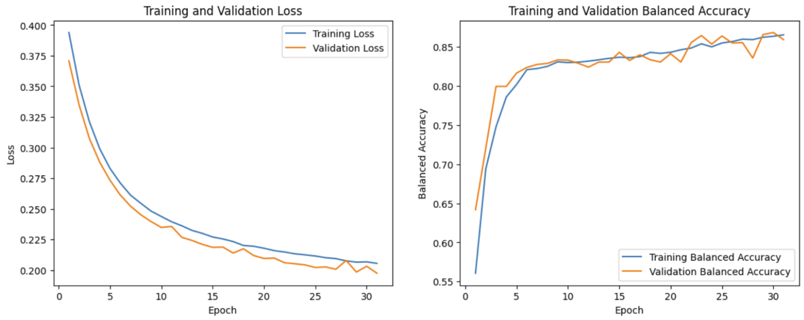

Figure 16: Loss and balanced accuracy plots for model 3

To summarize all the obtaiend results , we have the table

We see that the performance of the model decreased as test the model in the test set, proving that the model does not genralize well to unseen data

2.4. Limitations and Future Work

Although our models demonstrated good performance overall, there are several limitations that need to be addressed:

1-Small Dataset: The dataset's size may restrict our model's performance, having train samples from only 163 patients (the rest was for testing), and noting that the number of images per patient is not fixed. Increasing dataset size could enhance model generalization.

2-Class Imbalance: Imbalanced classes in the dataset might bias model predictions. Techniques like oversampling could help alleviate this issue

3-Data Quality: The quality of input data, including image resolution, could impact model accuracy. Ensuring high-quality data collection is crucial.

4-Generalizability: The model's performance may vary across different populations and healthcare settings. External validation is needed for broader applicability.

Future Works: To overcome these limitations we suggest:

1-Dataset Expansion: Collect more diverse data to improve model generalization.

2-Advanced Architectures: Explore advanced neural network architectures for better performance.

3-improve Augmentation: Design data augmentation techniques specific to images.

4-Clinical Validation: Conduct clinical validation studies for real-world applicability.

5-Interpret ability: Develop methods for interpreting model predictions to enhance trust.

3. Discussion

This study successfully applies deep learning to sort out lymphocytosis, reaching an impressive accuracy of **85.714%**. We experimented with a variety of models, starting from basic CNNs to the more complex ViT, highlighting our journey of fine-tuning and enhancements. We encountered several challenges along the way, such as ensuring our models performed well on new, unseen data, especially with the more advanced models. We also navigated the tricky balance of managing the dataset's imbalances and size without falling into the trap of overfitting. Interestingly, we discovered that our sophisticated ViT model didn't necessarily outperform the simpler ones. This experience underscored for us the importance of selecting and refining our models carefully to achieve both high accuracy and broad applicability.

4. Conclusion

To conclude, our journey in automating lymphocytosis classification showed us the power of deep learning. Using blood smear images and patient details from 204 people, we aimed to tell apart reactive from tumoral lymphocytosis. After lots of trial and error with different models, we acheived a balanced accuracy score of at **85.714%**. This highlights how our deep learning approach could make diagnosing faster and more accurate, really helping doctors make better decisions in treating blood disorders.

References

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.