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, poses diagnostic challenges. Our project seeks to automate classification using deep learning and a dataset of 204 individuals. By distinguishing between reactive and tumoral lymphocytosis, we aim to enhance diagnostic accuracy and streamline patient referral, improving hematology practice.

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

1. Introduction

Lymphocytosis, characterized by elevated lymphocyte count, presents diagnostic hurdles in healthcare. Manual blood smear examination, prone to subjectivity, drives our quest for automation through machine learning. Leveraging Lyon Sud University Hospital's dataset, we seek to distinguish reactive from tumoral lymphocytosis, augmenting diagnostic accuracy and aiding clinical assessment.

2. Methodology and Data

2.0.1. 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.

2.0.2. 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**

2.0.3. Data Visualization and analysis

Class Distribution: The visualization shows class '1' with 113 patients and class '0' with 50 patients. Age Distribution: The histogram suggests a quasi-normal distribution, predominantly between ages 60 and 80. **Age by Class Label:** Median age for class '0' is lower than class '1', with class '1' having a wider age range. **Age by Gender:** Median age is consistent between genders, with females showing a slightly wider range. **Gender Distribution:** Nearly equal numbers of males (82) and females (81) are observed.

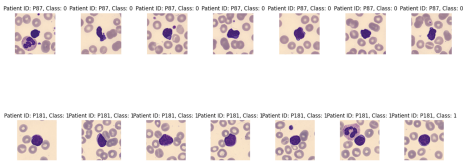

Figure 1: some sample images form class 0 and class 1 respectively.

2.0.4. Data Preprocessing

Class Balancing: lass Balancing: Figure 5 shows an imbalance in image distribution per patient, ranging from 16 to 198 images. The dataset also exhibits significant class imbalance, with the majority class (Label 1) outnumbering the minority class (Label 0). To address this, we employed two strategies: upsampling the minority class and downsampling the majority class. However, downsampling led to a loss of information and decreased model performance. Overall, these preprocessing steps mitigate bias towards the majority class, improving model robustness and generalization. See Figures 2, 3, and 4 for illustrations of oversampling and undersampling techniques.

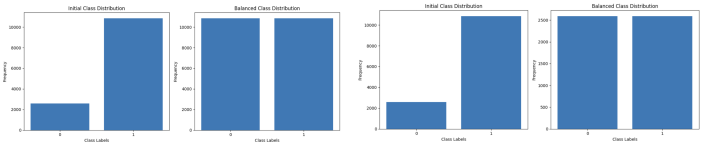

Figure 2: Oversampling the minority class

Figure 3: Undersampling the majority class

Figure 4: Balanced class distributions after oversampling and undersampling.

Data Augmentation:

We applied diverse data augmentation techniques, including resizing images, flipping, adjusting brightness, rotation, and applying transformations. By diversifying the training dataset, we aimed to prevent overfitting and improve the model's ability to generalize to unseen data, resulting in enhanced classification performance and clinical utility.

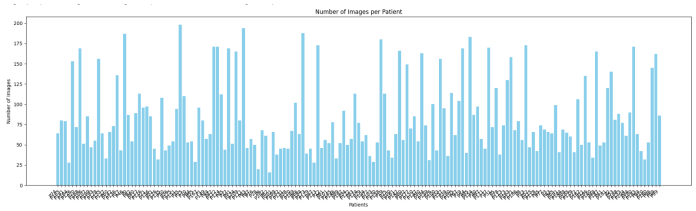

Figure 5: there is a clear imbalance between how the images are distributed per patiens

2.1. 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:

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

2.2. Experiments and results

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

2.2.1. First model

The initial model is a CNN designed for binary classification. It consists of three convolutional layers for feature extraction, followed by activation and max-pooling operations. Two fully connected layers and a sigmoid layer handle classification. Trained with AdamW optimizer (LR=0.01) and Binary Cross Entropy loss, it achieved a validation balanced accuracy of **85%**. See Figure 6 for loss and accuracy plots.

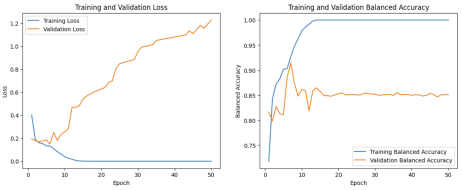

Figure 6: Loss and balanced accuracy plots for model 1

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 figure7.

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 .

2.2.2. Second Model

We improved performance by introducing a more complex CNN architecture with additional convolutional layers and dropout regularization to prevent overfitting. This model was trained using augmentation techniques outlined in section 2.0.4. Achieving **90%** validation accuracy, as shown in Figure 9, we observed significant instability at epoch 17, marked by a notable increase in validation loss and a corresponding decrease in validation score.

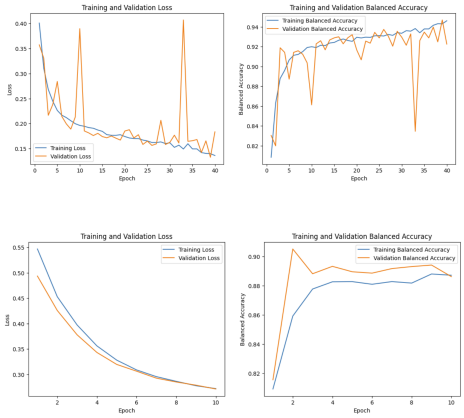

Figure 7: Loss and balanced accuracy plots for model 1 with balanced data.

2.2.3. 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 9.

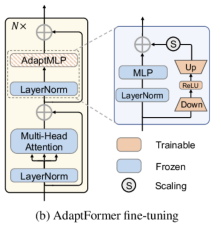

Figure 8: AdaptFormer

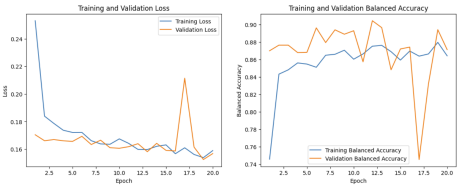

Figure 9: Loss and balanced accuracy plots for model 2.

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 10

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

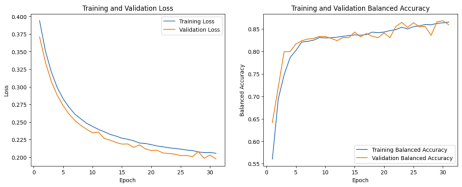

Figure 10: Loss and balanced accuracy plots for model 3

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

2.3. Limitations and Future Work

Our models demonstrated strong performance but face limitations:

Small dataset: With only 163 patient samples for training, increasing data volume is necessary. Class imbalance: Imbalanced classes may lead to biased predictions; oversampling techniques could address this issue. Data quality: Ensuring high-quality data, particularly regarding image resolution, is paramount. Generalizability: Validating model performance across diverse populations is crucial. Future work includes: Dataset expansion: Collecting a more diverse range of data. Advanced architectures: Exploring improved neural network designs. Improved augmentation: Developing customized image augmentation techniques. Clinical validation: Validating models in real-world clinical environments. Interpretability: Enhancing trust through interpretable model predictions.

3. Discussion

This study applied deep learning to classify lymphocytosis with **85.714%** accuracy. Experimentation with different models revealed challenges such as dataset imbalances and non generalizability of the models. Surprisingly, simpler models performed comparably to complex ones, emphasizing the importance of careful model selection for accuracy and broad applicability.

4. Conclusion

In conclusion, our deep learning approach aimed to automate lymphocytosis classification using data from 204 individuals. After experimenting with various models, we achieved a balanced accuracy of **85.714%**. This highlights the potential of deep learning to enhance diagnostic accuracy and assist clinicians in treating blood disorders more effectively.

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.