Classification of Lymphocytosis from Blood Cells

Aissa ABDELAZIZ and **Valentin Gerard**
Team : Infusion.ai

Abstract

In this report, we provide a comprehensive explanation of our approach to tackling the Kaggle Challenge on the Classification of Lymphocytosis from Blood Cells.In Section 2, we provide the methodologies utilized, focusing on both the approaches adopted and the model developed. In Section 3, we explain the various strategies implemented to enhance the performance and score of our model. Our aim is to optimize the effectiveness of the proposed model by utilizing a wide array of techniques.

Code source

1. Introduction

Lymphocytosis, characterized by an elevated lymphocyte count in the bloodstream, can stem from various causes, ranging from reactive responses to infections or stress to more serious underlying lymphoproliferative disorders, indicative of malignancy.

The best way for determining whether lymphocytosis is malignant is flow cytometry, yet its expense and time-consuming nature limit its widespread use for all patients. Consequently, a preliminary diagnostic approach, utilizing clinical attributes like age and lymphocyte count alongside data extracted from blood cell images, such as lymphocyte size and texture, is usually employed.

we try to propose a solution for automated classification of lymphocytosis (tumoral or not) based on medical information of the patient and from the images of the patient's blood cells .

The main difficulties of this challenges are to find an efficient way to handle these multimodal data and also a way to address multiple instances. Indeed, the annotations provided for the images is at the patient scale and we do not have a specific label for each image.

2. Architecture and methodological components

2.1. Dataset

The dataset is made up of a collection of blood smears and patient attributes (age, gender and lymphocyte count) and was undertaken at the routine hematology laboratory of Lyon Sud University Hospital. Only symptomatic patients with lymphocyte count exceeding $410^9/L$ are considered. Of the 204 subjects in the dataset, 142 were allocated for training purposes, with a distribution of 44 cases classified as reactive lymphocytosis and 98 as malignant lymphoproliferative disorders.

As a first step, we analyze the dataset based on its attributes. We observe that the Gender attribute is balanced between the positive and negative classes. Regarding the

age attribute, we note that patients under the age of 50 are less likely to test positive. Additionally, examining the lymphocyte count attribute, which represents the quantity of lymphocyte cells in the patient's blood, we find that higher values correlate with a higher probability of being affected (positive). These observations are visually represented in Figure 1.

Finally, we decide to quantize the gender attribute by mapping Male to the value 1 and Female to the value 0. Additionally, we generate an "Age" attribute based on the Date of Birth. This step is necessary to develop a classification model that utilizes numerical attributes.

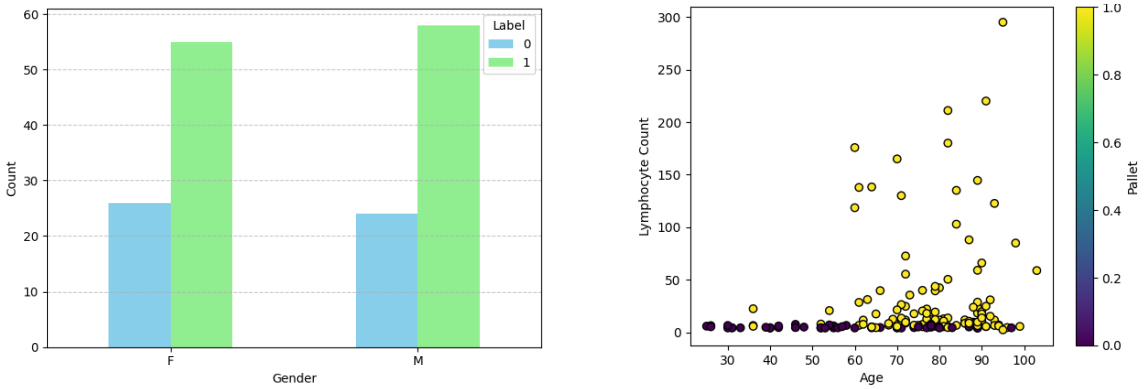

Figure 1: Combined analysis of gender distribution and age-lymphocyte correlation with respect to labels.

2.2. Baseline Model: Support Vector Machine (SVM) Classifier

After analyzing the attributes of the provided dataset: "Gender,"Age","lymphocyte count", our initial intuition is to utilize this information to predict the class using a Support Vector Machine (SVM) classifier (Tang, 2015). This choice is motivated by its ability to handle unbalanced data, good generalization performance, and robustness to overfitting. As a result, we achieved a balanced score of 79.8% on the test set.

2.3. Developed Model: Enhanced Classification Approach

In order to enhance our prediction accuracy, we have decided to use the images provided within the dataset. These images contain valuable information regarding the blood smears of the patients, which can significantly aid our classifier in accurately predicting the affected individuals.

To extract pertinent features from these images, we use various pre-trained models including ResNet18, ResNet34, ResNet50 (Kaiming He, 2015), and VGG-16 (Karen Simonyan, 2014). These models are renowned for their effectiveness in feature extraction tasks, enabling us to leverage a diverse range of learned features to enrich our classification process.

Following preprocessing, which includes normalizing the images, we perform feature extraction on all the images provided for each patient. We concatenate these features with the attributes "Gender," "Age," and "Lymphocyte Count" to create a feature vector for each subject. Subsequently, we train classifier models such as MLP and XGBoost(Tianqi Chen, 2016) to make predictions using these features. Our proposed framework was inspired by the paper by (Mihir Sahasrabudhe, 2020), and it is described in detail in Figure 2.

Due to the dataset's imbalance between the two labels, the initial intuition was to utilize weighted loss. In our case, we employed Binary Cross-Entropy loss, which is commonly used in binary classification tasks, and assigned more importance to the minority class (label 0). Consequently, we attained a balanced score of 85.24% on the test set. We will delve into further detail regarding other approaches we utilized to tackle the challenge of unbalanced data at the end of the next sections.

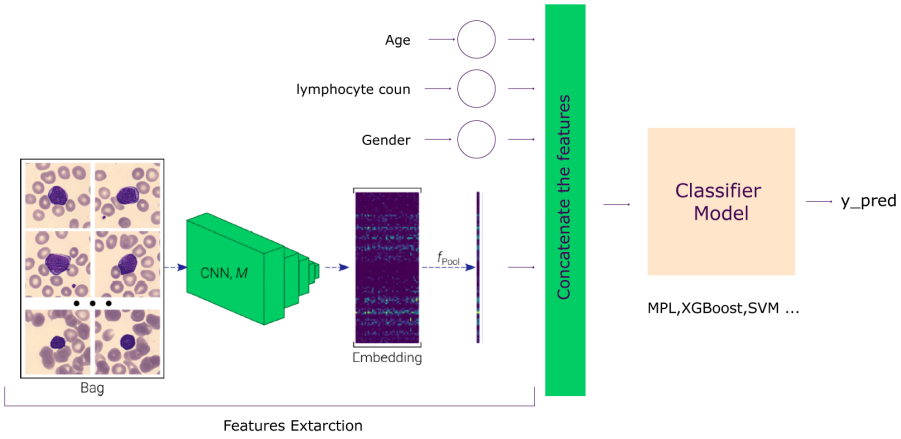

Figure 2: Developed Model: Enhanced Classification Approach

3. Model tuning and comparison

3.1. Validation procedure

In our training process, we split our dataset into 80% for training and 20% for validation sets, respectively. Furthermore, to ensure the robustness of our models, we employed a 3-fold cross-validation approach. This method allowed us to assess the consistency of our models across different subsets of data.

As our dataset is strongly unbalanced, we choose adapted metrics to compare our model such as balanced accuracy and F1-score that are relevant in this unbalanced scenario. Moreover, as a medical application we do not want to miss a positive patient thus we decided also to consider the recall for the evaluation of our models.

3.2. Training parameters

In order to compare our models under the same conditions, we adopt the same training parameters described in Table 1. We employ the Adam optimizer, which dynamically ad-

justs learning rates for individual parameters by utilizing past gradients, thereby efficiently optimizing model parameters. This method combines the advantages of momentum and RMSProp. Furthermore, we utilize a batch size of 8 to improve both generalization performance and convergence properties of the optimization algorithm.

Table 1: Training parameters

Parameter	**Value**
Optimizer	Adam
Learning Rate	0.001
Batch Size	8
Number of Epochs	100

3.3. Ablation study

Table 2: Comparison of Dice Score for Different Loss Functions

Metrics on different validation folds									
	Balanced Accuracy			F1 Score			Recall		
Architecture	1	2	3	1	2	3	1	2	3
SVM with medical information	0.82	0.81	0.80	0.87	0.86	0.87	0.85	0.83	0.85
SVM with med. inf. and ResNet features	0.78	0.9	0.58	0.87	0.89	0.83	0.87	0.8	0.96
SVM with ResNet features	0.57	0.61	0.60	0.23	0.87	0.53	0.13	0.96	0.39
MLP with med. inf. and ResNet features	0.91	0.86	0.89	0.90	0.84	0.88	0.83	0.72	0.78
MLP with ResNet features	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0

First, we can notice that our baseline, using only the medical information about the patient (Age, Gender and Number of Lymphocytes) demonstrates a strong consistency over the validation folds for each metrics. By giving to the SVM also the image features extracted by ResNet, we lost in performances and the consistency over the folds. It seems that the structure of the extracted features is too complicated for the SVM and justify our choice to use a MLP to extract relevant information from this structure.Indeed, the MLP outperforms significantly our baseline by conserving the consistency over the folds. Thus it is the model that we use for the challenge with a precision of **83.13** % on the test set. However, it important to notice that its recall is around 80 % This means that for 1 out of 5 patients with a malignant case, we will not recommend undergoing flow cytometry .

Finally, by comparing it to the MLP model that use only the extracted features for which we get the worst performances possible, we can say that there are the medical information

which are simple pieces of information which guide the model to connect the complex extracted features to the classification task.

3.4. Unsuccessful ideas

To handle the difficulty of multiple instance learning our first idea was to use as input an image which was the combination of several instances, the intuition behind it was that by having access to a sufficient number of instances we could limit for positive patients the influence on the predictions of the few instances containing normal lymphocytes. We try as input an image with 10 channels where each channel is an instance of the patient. However, by doing this we lost the spatial alignment between the channels and the possibility to use pre-train vision model that are usually train for 1 or 3 channels input. Then, we also tried to use as input a grid of 9 images but by doing this we lost in resolution. With both approaches our balanced accuracy was always under 0.60 and the pooling strategy over the instances described before was more efficient.

Additionally, we attempted to balance the dataset by generating new data through various techniques, including random rotation, horizontal and vertical flips, and histogram shifting. Subsequently, we utilized a pre-trained model to extract features, and employed SMOTE to generate new samples for attributes such as age, gender, and lymphocyte count, which were associated with the newly generated features. Through these methods, we successfully balanced and augmented the dataset. However, we did not observe any improvements in the obtained results. This could be attributed to the mismatch between the generated data and the provided dataset, suggesting that the manner in which we generated the data may not align with the characteristics of the provided data.

3.5. Future Work

As part of our future work, we propose to explore alternative approaches for data augmentation to ensure alignment with the original data, thereby aiding the model in effectively generalizing to unseen data. Additionally, we suggest developing models such as VAE trained on the provided data for tasks such as image generation, utilizing the encoder segment of these models as our feature extractor.

Furthermore, we propose implementing ensemble learning techniques across various models, combining different feature extractors and classifier models to enhance overall performance scores.

3.6. Conclusion

We proposed a solution based on a MLP combining the features extracted by ResNet from the images to the age, the gender and the concentration of lymphocyts. We observed how these medical information are important for the model in order to connect the features extracted by ResNet to the classification task. Our solution addressed the difficulties of multiple instance learning by pooling the features extracted from each images and reach good performance. However, it is important to keep in mind that without a recall closer of 1 our solution is not suitable for a direct application for diagnostic of Lymphocitosis where we absolutly want to avoid false negative.

References

Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. Eprint arXiv:1512.03385, 2015.

Andrew Zisserman Karen Simonyan. Very deep convolutional networks for large-scale image recognition. Eprint arXiv:1409.1556, 2014.

Evangelia I Zacharaki Eugénie Maurin Béatrice Grange Laurent Jallades Nikos Paragios Maria Vakalopoulou Mihir Sahasrabudhe, Pierre Sujobert. Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis. Eprint hal-03032875, 2020.

Yichuan Tang. Deep learning using linear support vector machines. Eprint arXiv:1306.0239 , 2015.

Carlos Guestrin Tianqi Chen. Xgboost: A scalable tree boosting system. Eprint arXiv:1603.02754, 2016.