Lymphoma Subtype Classification Using Neural Networks To Support Human Hematopathologist decisions

Chiara Fantinato†, Riccardo Malgarotto Boschiero†

***Abstract*—Lymphoma classification is a complex task even for expert hematopathologists. Therefore, the implementation of algorithms for the automated classification of lymphoma subtypes could be helpful in supporting physicians' decisions. In this paper, we exploited neural networks based on residual units to distinguish between the three most common lymphoma subtypes (CLL, FL, MCL) consolidating that splitting images into patches is a winning strategy and demonstrating that good performances are achievable also by decreasing model complexity. We also demonstated that the management of artifacts impact on classification performances by including in our pre-processing pipeline a strategy to clean images. Results shows that the high degree of staining variation is not a limit for the classifiers, because including in our processing pipeline Macenko normalization does not improve performances. In addition, we provided visual explanations for machine' s decisions through the computation of heatmaps, to make the process more transparent and create a meeting point between machine learning and clinicians. A further development of the current study could be to ask to human hematomathologists to annotate regions of interest in histological images to verify the clinical validity of our heatmaps.**

***Index Terms*—Subtype Lymphoma Classification, Supervised Learning, Neural Networks, Residual Units, Heatmap, Visual Attention, Exlainable Deep Learning.**

I. Introduction

Lymphoma is a type of cancer that begins in lymphocytes, which are the infection-fighting cells of the human body. The delivered treatment depends on the type and stage of the lymphoma. For this reason, it is necessary for the doctor to make an accurate diagnosis [1]. Currently, the gold standard for lymphoma subtypes identification is based on morphologic features of the tumor as observed by light microscopy of hematoxylin (H)- and eosin (E)-stained tissue sections and interpreted by an expert hematopathologist [2], although the 5th edition of the World Healt Organization Classification of Haematolymphoid Tumours (WHOHAEM5) recognizes the increasing importance of genetic and other molecular data in the evaluation of lymphoid neoplasia [3].

Lymphoma classification is a complex task even for expert hematopathologists, due to numerous factors such as the variation in slide staining and sectioning. For this reason, automated classification of Lymphoma subtypes is appealing in supporting physicians' decisions. The implementation of an accurate lymphoma classifier is challenging because of several uncertainties. Firstly, histological lymphoma features could be focused on a restricted area of the slide. Secondly, a single magnification may be insufficient to distinguish between different types of lymphoma. Finally, a tumor might contain a range of cell types that, although derived from the same clone, do not share cytologic characteristics [2]. Several machine learning approaches have been proposed over the last years to construct computer-aided diagnostic (CAD) systems. [2] demonstrated that computer vision can be successful in discriminating between three of the most common lymphoma types. [4], [5], [6], [7], [8], [9] showed the potential of deep learning, a sub-domain of machine learning, in the contribution to the automation of diagnosis. Specifically, [8] and [9] adopted the so-called explainable deep learning approach to tackle one of the main limitations in showing impact of machine learning methods: the lack of transparency about the reasoning behind machine's decisions.

In this paper, we classified Chronic Lymphocytic Leukemia (CLL), Follicular Lymphoma (FL), and Mantle Cell Lymphoma (MCL) presenting two neural network-based architectures: a ResNet50-based architecture and a ResNet built from scratch. To train our models we used a NIA curated dataset including samples prepared by different pathologists at different sites. There is a large degree of staining and sectioning variation and this is an attempt to reproduce the training of human pathologists in a clinical setting. We inclueded in our pre-processing pipeline a strategy to deal with artifacts improving results. We demonstrated that staining variation is not a limit for our classifier, because including in our pipeline the method for normalizing histological slides proposed by Macenko et at [10] does not improve performances. Furthermore, we achieved classification performances comparable to what can be found in literature, even by decreasing the complexity of the exploited model and consequently the computational efforts required. Moreover, we were able to produce visual explanations for machine's decisions to make the process more transparent.

This report is structured as follows. In Section II we describe the state of the art. The high level description of our architectures is reported in section III, while in section IV we describe the analyzed signals. The proposed signal processing technique is detailed in Section V and its performance evaluation is carried out in Section VI. Concluding remarks are

†Second cycle degree in Bioengineering, email: chiara.fantinato.4@studenti.unipd.it

†Second cycle degree in Bioengineering, email: riccardo.malgarottoboschiero@studenti.unipd.it

provided in Section VII.

II. Related Work

Several methods based on the dataset we used for the current paper have been proposed. In 2010, Orlov et al. [2] demonstrated that whole-image pattern recognition without reliance on segmentation can be successful in discriminating between CLL, FL, and MCL, with a classification accuracy of 98%. The authors considered gray, RGB, Lab, and HE color spaces, and analyzed transform-based global features from these color spaces using three different classifiers: weighted-neighbor distance, radial basis functions and naive bayes classifier. Fisher scores, collinearity measures, and Pearson correlations were used to rank the features. Fisher linear discriminant, minimum redundancy maximum relevance, and Fisher/Correlation algorithms were tested to select features. Significant differences in classification accuracies were found only between the four color spaces tested, with HE consistently more accurate than the others followed by RGB, gray, and Lab. Different classifiers lead to comparable results, as well as for different feature selection algorithms. In 2016, Andrew Janowczyk and Anant Madabhushi [4] split the images in the dataset into patches and used them to classify lymphoma subtypes through AlexNet architecture. They obtained 96.58% accuracy by assigning to an image the class with the highest number of votes. The authors noticed that misclassification of images correlated with the presence of some type of artifact created during the scanning process. Starting from this observation, we implemented a strategy to clear images from artifacts and to improve classification performance. In 2019, Tambe et al. [5] proposed a method to classify lymphoma subtypes based on the Inception V3 network, which yields 97.33% accuracy. In 2021, Zhang et al. [6] used ResNet50 to get 98.63% accuracy. In 2022, Hathem et al. [7] divided the images in the dataset into patches to perform data augmentation, and achieved 98.7% accuracy using an optimization of the CNN algorithm. We referred to this as our benchmark. [4], [5], [6], [7] preferred transfer learning over full learning to be less prone to overfit, which means that they fine-tuned a pre-trained network.

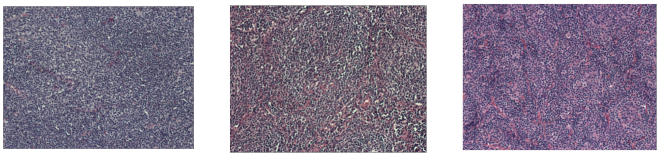
We believe that all the aforementioned works have two common limitations. Firstly, they are based on the assumption that each part of the image contributes equally in classifying lymphoma subtypes. Mimicking the way human hematopathologists make decisions, by Introducing visual attention mechanisms into the classification process to focus the attention on restricted areas of the images, could be helpful for improving performances of the machine. Secondly, they do not provide insights about the way outputs are produced in response to given inputs. Especially in the medical domain, not justifying how a decision is made could be a bottleneck in showing impact of machine learning models, as it is difficult for clinicians to trust the process. In 2023, Ammar et al. [8] proposed a method to distinguish between lymphoma, carcinoma, and benign lesion. The approach splits the images into patches, classifies them using the InceptionResNetV2 model, and takes the average prediction score over all the patches of an image as prediction result. To help the pathologist further, the authors also provided an activation heatmap for each patch, using Gradient-weighted Class Activation Map (Grad-CAM) [11]. The activation heatmap highlights the important regions in the related patch. Inspired by this approach, we also retrieve activation maps for our images using the same method to provide to doctors not only a decision but also a visual explanation for that decision. In 2018, Guan et al [9] dealt with the task of thorax disease classification on chest X-ray images proposing a three-branch attention guided convolutional neural network (AG-CNN). The global branch consists of a variant of ResNet50 and is trained to classify based on the entire images. The local branch is also based on ResNet50 design, but learns to classify based on cropped images, obtained using the attention heatmap generated from the global branch. The fusion branch computes the final class by concatenating the last pooling layers of both the global and the local branches.

III. Processing Pipeline

Starting from the NIA dataset composed of histological slides of the three most common lymphoma subtypes (CLL, FL, MCL), we exploited neural network architectures to distinguish between the represented classes. The dataset is randomly divided into training, validation, and test set according to 6:2:2. Training set is used for learning; validation set is used for tuning hyper-parameters and model optimization; test set is used to assess generalization ability of models. The dataset is pre-processed using a gaussian filter to remove possible artifacts, histogram equalization to enhance contrast, Macenko normalization to reduce the high degree of staining variation. It is also augmented by dividing images into patches or rotating and/or flipping them to avoid overfitting. When images are divided into patches, when images are divided into patches, the probability of belonging to a class is obtained by summing the output probabilities of each patch. When images are not divided into patches, they are resized to lower dimensions for reducing computational efforts. We can divide our modeling process into two parts. Firstly, we exploited already existing models (ResNet50). Secondly, we built a ResNet from scratch to investigate if comparable performances were possible by decreasing model complexity.

A. Transfer learning with ResNet50

We started our work using transfer learning, by fine-tuning on our datasets pre-trained models on Imagenet data from Keras. In Particular, we based this first part of our modeling process on ResNet50. At the very beginning, we simply exploited the ResNet50 architecture, without apporting any structural changes. At a later stage, we decided to retrieve visual explanations for machine's decisions, to make the process more transparent, through the computation of heatmaps using Grad-CAM. Then, inspired by the approach explained in [9], we exploited these heatmaps to investigate the importance of focusing attention on restricted parts of the images to



(a) Chronic Lymphocytic Leukemia (CLL). (b) Follicular Lymphoma (FL). (c) Mantle Cell Lymphoma (MCL).

Fig. 1: Samples for each class.

Dataset	CLL	FL	MCL	Total
Training	68	86	70	224
Validation	24	26	25	75
Test	21	27	27	75

TABLE 1: Datasets compositions

improve classification performances by creating an attention-guided ResNet50 (AG-ResNet50). Specifically, we treated our previous ResNet50 model as the global branch, aimed to classify based on the whole images, and we created another pre-trained ResNet50 model on ImageNet to be used as the local branch, aimed to classify by focusing its attention on restricted areas of the images. We fed the local branch using masked images obtained applying a threshold to the heatmaps generated from the global branch. The fusion branch computes the final class by concatenating the last pooling layers of both the global and the local branches.

B. ResNet built from scratch

In the second part of our modeling process we investigate the ability of less complex models in classifying lymphoma subtypes. Firstly, we created a ResNet model with few layers and we tested its classification ability on our dataset. Secondly, based on the fact that histological lymphoma features could be focused on a restricted area of the slide, we included in our architecture another ResNet model as local branch and, as for the AG-ResNet50, we fed it by using masked images obtained from global branch heatmaps.

IV. Signals

The NIA curated dataset used in this study consists of 374 images of size 1388x1040. There are 113 images for Chronic Lymphocytic Leukemia (CLL), 139 images for Follicular Lymphoma (FL) and 122 images Mantle Cell Lymphoma (MCL). The dataset includes samples prepared by different pathologists at different sites, resulting in a large degree of staining and sectioning variation. A sample for each class is shown in Fig. 1.

We split the dataset into training, validation, and test sets, made of 60%, 20%, and 20% images respectively. The compositions of the training, validation, and test sets are shown in TABLE 1.

Starting from the observation of a pattern between misclassified images and presence of some type of artifacts created during the scanning process in [4], we thought that poor quality of slides could impact on model performances and we decided to remove artifacts from affected images using a 3x3 gaussian filter (Fig. 2). Images affected by artifacts were detected automatically by considering the variance of images filtered through a Laplacian filter: variance values greater than a certain threshold corresponds to presence of artifacts. Considering the training set, we observed that 700 was a robust threshold.

Our pre-processing pipeline includes also histogram equalization, to stretch out the intensity range and improve image contrast, and/or Macenko normalization [10], to reduce the high degree of staining variation by performing intensity correction through the standardization of the level of chemical concentration of H and E (Fig. 3).

Due to the restricted dimension of the dataset, we augmented the dataset by splitting images into patches or by rotating and/or flipping them. Patches were created in two ways using the training set: by resizing the images from 1388x1040 to 1300x1040 and splitting each of them into 20 non-overlapping patches of 260x260 with a stride of 260; by resizing the images from 1388x1040 to 1300x1040 and splitting each of them into 63 overlapping patches of 260x260 with a stride of 130. When images were not divided into patches, they were resized from 1388x1040 to 224x224. Through transformations, 8 rotated and flipped versions of each image were retrieved (Fig. 4).

V. Learning Framework

In this section we are going to explain how the learning process is carried out. We are dealing with a multi-class classification problem, where each input image $\mathbf{x}$ has associated a label $\mathbf{t}$:

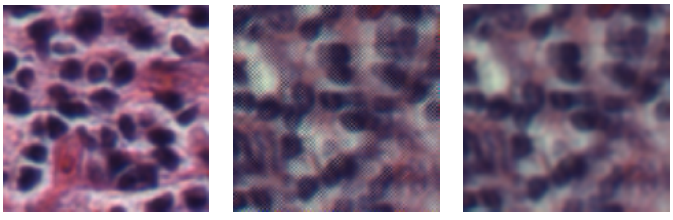
$$\mathbf{t} = [t_1, ..., t_K]^T with : t_i \in 0, 1, \sum_i t_i = 1 \qquad (1)$$

A. Error Function

We opted for categorical cross-entropy as error function:

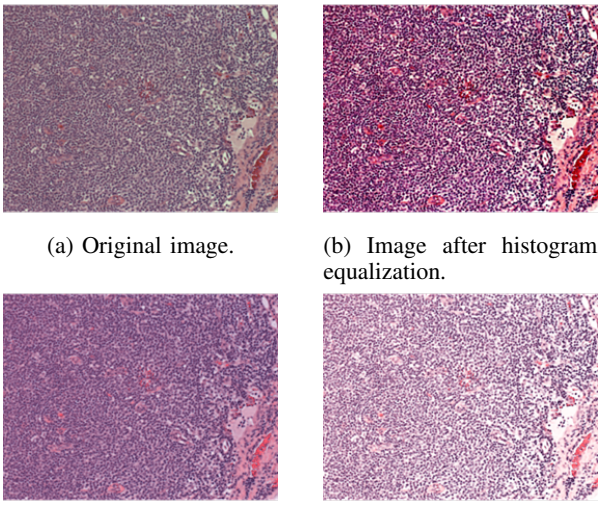
$$E(\mathbf{x}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \log(y_k(\mathbf{x_n}, \mathbf{w})). \qquad (2)$$

where $y_k(\mathbf{x}, \mathbf{w}) = p(t_k = 1|\mathbf{x})$ is the network output from neuron k for input x_n.



(a) Image without artifacts. (b) Image affected by artifacts. (c) Filtered image.

Fig. 2: Artifacts managment.



(a) Original image.

(b) Image after histogram equalization.

(c) Image after Macenko normalization.

(d) Image after histogram equalization and Macenko normalization.

Fig. 3: Variation management.

B. Output Function

We opted for softmax function as output function, which returns as output values the probability that the input x belongs to class k. The output at node k for input x is computed as:

$$y_k(\mathbf{x}, \mathbf{w}) = \frac{\exp(z_k(\mathbf{x}, \mathbf{w}))}{\sum_j \exp(y_j(\mathbf{x}, \mathbf{w}))}. \quad (3)$$

where z_i are the output of the last hidden layer.

C. ResNets

ResNets consist of several stacked convolutional and residual blocks (Fig. 5). Each residual block can be expressed as [12]:

$$\mathbf{y_l} = h(\mathbf{x_l}) + \mathcal{F}(\mathbf{x_l}, \mathcal{W_l}). \quad (4)$$

$$\mathbf{x_{l+1}} = f(\mathbf{y_l}). \quad (5)$$

where $\mathbf{x_l}$ and $\mathbf{x_{l+1}}$ are input and output of the $i-th$ unit, $\mathcal{F}(\cdot)$ is a residual mapping, $h(\cdot)$ is an identity mapping, $f(\cdot)$ is a ReLU function. During the training of ResNet50, weights of the network are initialized to the weights computed on ImageNet and then adapted to our dataset. AdaMax optimizer is used and fixed and variable learning rates are tested.

D. Proposed ResNet built from scratch

The architecture of the proposed ResNet is summarized in Fig. 6. AdaMax optimizer is used and fixed and variable learning rates are tested.

E. Heatmap and masked images

Heatmaps were extracted from a weighted average of the last convolution layers using as weights the parameters the last dense layer as suggested in [11]. Masks were created from heatmaps by scaling them between 0 and 1 and setting a threshold at 0.6. Morphological operators (opening and dilation) were exploited to eliminate small useless areas and to expand a bit the interested region. Fig. 7 shows examples of heatmaps, related masks and masked images for both ResNet50 and the proposed ResNet. Since clinical indications were not available, we decided to evaluate the validity of the areas of interest indicated by the heatmaps as follows. When we performed data augmentation by rotating and flipping images we got 8 versions of each. Since it is reasonable to think that if the network works well it must recognize consistent important areas among different versions of the same image, we evaluate the intersection of the masks obtained from the heatmaps of these images. The intersection image was created setting to one the pixels identified as important by at least 4 masked images out of 8. Particularly, we computed for each image the ratio between the number of non-zero pixels in the intersection image and the number of non-zero pixels in the image obtained by summing all the masked images related to that image, and we compute mean. We also retrieved the number of images without intersection. These metrics are evaluated on the 224x224 version of the dataset for computational reasons.

VI. Results

A. Classification Performances

The optimization process of the architecture based on ResNet50 results in 10 epochs, 32 batch size, AdaMax optimizer, and variable learning rate from 0.001 to 0.0001 as the best setup.

Using the optimized architecture based on ResNet50, we investigated the classification performances applying different pre-processing and data augmentation combinations. Results (TABLE 4) shows that: best validation accuracy is obtained by removing artifacts and augmenting the dataset using patch approach; differences in validation performances splitting images into non-overlapping or overlapping patches is too low to worth the increase in computational efforts required in the case of overlapping patches; applying Macenko normalization worsen validation accuracy, suggesting that the high degree of staining variation is not a limitation for the model. Training and validation loss and accuracy versus number of epochs using ResNet50 on the dataset pre-processed to remove artifacts and augmented by splitting images into non-overlapping patches are shown in Fig. 8.

The optimization process of the proposed ResNet results in 15 epochs, 32 batch size, AdaMax optimizer, and variable learning rate from 0.001 to 0.0001 as the best setup (TABLE 5). Training and validation loss and accuracy versus number of epochs using the proposed ResNet on the dataset pre-processed to remove artifacts and augmented by splitting images into non-overlapping patches are shown in Fig. 9.

Finally, we applied the final models on the test set to evaluate their ability to generalize. Results (TABLE 6) shows that performances of our models are comparable to previous works found in literature, with the proposed ResNet classifying slightly worse. For the proposed ResNet we computed: the confusion matrix (Fig. 10); TPR, TNR, FPR, and precision for

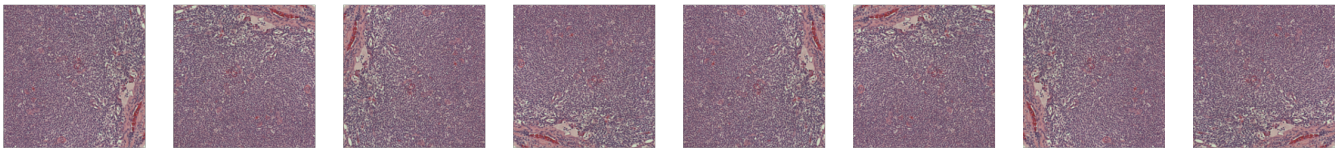


Fig. 4: Transformation on an image.

Pre-process	**Data augmentation**	**Training accuracy [%]**	**Validation accuracy [%]**
No pre-process	Rotating and flipping	99.67	97.33
Removing artifacts	Rotating and flipping	99.72	98.67
Removing artifacts, histogram equalization	Rotating and flipping	99.94	96.00
Removing artifacts, Macenko normalization	Rotating and flipping	99.94	92.00
Removing artifacts, histogram equalization, Macenko normalization	Rotating and flipping	100	89.33
No pre-process	Splitting into non-overlapping patches	99.89	98.60
Removing artifacts	Splitting into non-overlapping patches	99.98	**99.00**
Removing artifacts	Splitting into overlapping patches	99.99	**99.01**

TABLE 2: Classification performances for different pre-processing and data augmentation combinations using ResNet50.

Epochs	**Batch size**	**Steps per epoch**	**Learning rate**	**Training accuracy [%]**	**Validation accuracy [%]**
10	32	140	1e-3	97.10	94.00
15	32	140	1e-03 to 1e-04	98.48	**98.07**
25	32	140	1e-03 to 1e-07	99.08	96.80
10	64	70	1e-03 to 1e-04	98.64	97.27
25	64	70	1e-03 to 1e-07	98.12	96.00

TABLE 3: Classification performances of the proposed ResNet on training and validation sets pre-processed to remove artifacts and augmented by splitting each image into non-overlapping patches.

each class (TABLE 7); the F-measure as an alternative metric to accuracy (0.946). Our ResNet misclassifies 4 of 75 images of the test set.In particular, MCL class has TPR equal to 1, denoting that all its images have been correctly classified; CLL class has FPR equal to 0 and precision equal to 1, showing that all the images predicted as CLL are correctly classified.

In addition, we found that the introduction of visual attention in our architecture doesn't works on our dataset. As matter of fact the accuracy decreased adding to the model a local branches which classifies focusing on a restricted area of the images: for ResNet50, validation accuracy is unchanged at 98.67%; for the proposed ResNet, validation undergoes a deterioration going from 97.33% to 93.33%. An explanation for that could be that we fed the local branch with masked images obtained from both good and bad intersections (Fig. 11).

B. Validity Of Heatmaps

Computing for each image the ratio between the number of non-zero pixels in the intersection image and the number of non-zero pixels in the image obtained by summing all the masked images related to that image and taking the mean In a. Results (TABLE 9) shows that the average proportion of common pixels is higher for the heatmaps coming from

Model	Test accuracy [%]
Orlov et al. [2]	98.00
Andrew Janowczyk and Anant Madabhushi [4]	96.58
Tambe et al. [5]	97.33
Zhang et al. [7]	98.63
Hathem et al. [7]	98.67
ResNet50	98.67
Proposed ResNet	94.67

TABLE 4: Classification performances of ResNet50 and the proposed ResNet on test set pre-processed to remove artifacts and augmented by splitting each image into non-overlapping patches, compared to previous work found in literature.

Class	TPR	TNR	FPR	Precision
CLL	0.857	1	0	1
FL	0.963	0.957	0.043	0.929
MCL	1	0.957	0.043	0.931

TABLE 5: TPR, TNR, FPR, precision on test set for the proposed ResNet.

the proposed ResNet (0.4 versus 0.6). Furthermore, using our model there are no images without intersection, while using ResNet50-based architecture there are 10 images without intersection. Probably, the explanation for this is that the heatmaps for our ResNet are less selective than the ones related to ResNet50, which means that in the first case heatmaps suggest areas of image where the physicians should not focus on, while in the second case heatmaps suggest to doctors where to focus their attention.

C. Computational Efficiency

The training of ResNet50 on Colab last for the 14112 overlapping patches 90 min/epoch without GPU accelerator and 7 min/epoch with GPU; for the 4480 non-overlapping patches 23 min/epoch and 79 sec/epoch; for 1792 resized images 6 min/epoch and 20 sec/epoch. Given the time and computational resources required to train the ResNet50 and given the size of its 23,540,739 parameters (53,120 non-trainable) that occupies 270.4 MB, we decided to create our neural network from scratch with the purpose to reduce model complexity by keeping the accuracy as high as possible. Our proposed ResNet has 444,163 parameters (2,240 non-trainable) and occupies 5.4 MB, saving more than 98% of the previously occupied memory with a good predictive power. On the other hand, the training time on Colab is practically the same compared to ResNet50.

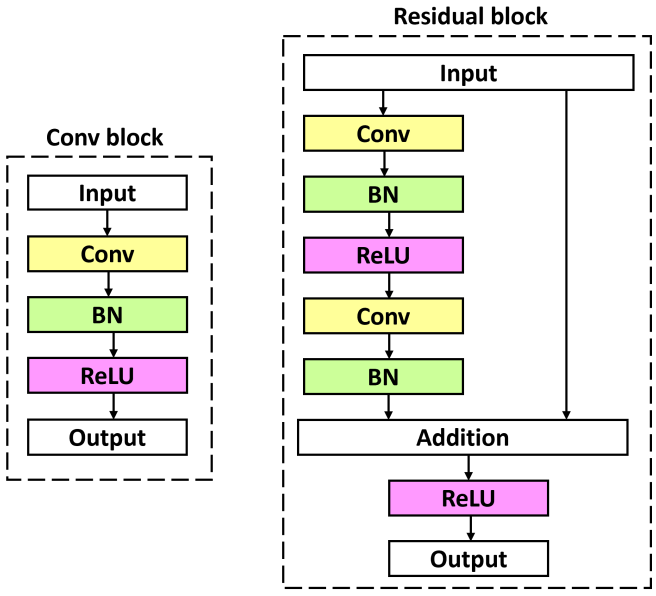


Fig. 5: Convolutional and residual blocks.

Model	Branch	Training accuracy [%]	Validation accuracy [%]
ResNet50	Global	99.72	98.67
	Local	100	98.67
	Fusion	100	98.67
Proposed ResNet	Global	98.00	97.33
	Local	99.00	83.00
	Fusion	99.83	93.33

TABLE 6: Training and validation accuracy for the architectures including visual attention.

Model	Mean [%]	Images without intersection
ResNet50	40	10
Proposed ResNet	60	0

TABLE 7: Metrics to evaluate quality of heatmaps.

VII. Concluding Remarks

In this work we investigated the ability of neural network-based models in classifying the three most common lymphoma subtypes (CLL, FL, MCL). The dataset was pre-processed using a gaussian filter to remove possible artifacts, histogram equalization to enhance contrast, Macenko normalization to reduce the high degree of staining variation. It was also augmented by dividing images into patches or rotating and/or flipping them to be less prone to overfitting. When images are divided into patches, the probability of belonging to a class is obtained by summing the output probabilities of each patch. We demonstrated that high degree of staining variation is not a limit for our classifier, because including in our pipeline the Macenko normalization does not improve performances. Then, classification performances slightly lower but comparable to what can be found in literature are possible even by decreasing the complexity of the exploited model and consequently the computational efforts required. Applying the proposed ResNet on the test set we get high values of TPR, which is good: since our aim is to identify lymphoma subtypes to take actions (e.g., to deliver treatment), we want that what we identify as positive would be really positive. Finally, introducing visual attention

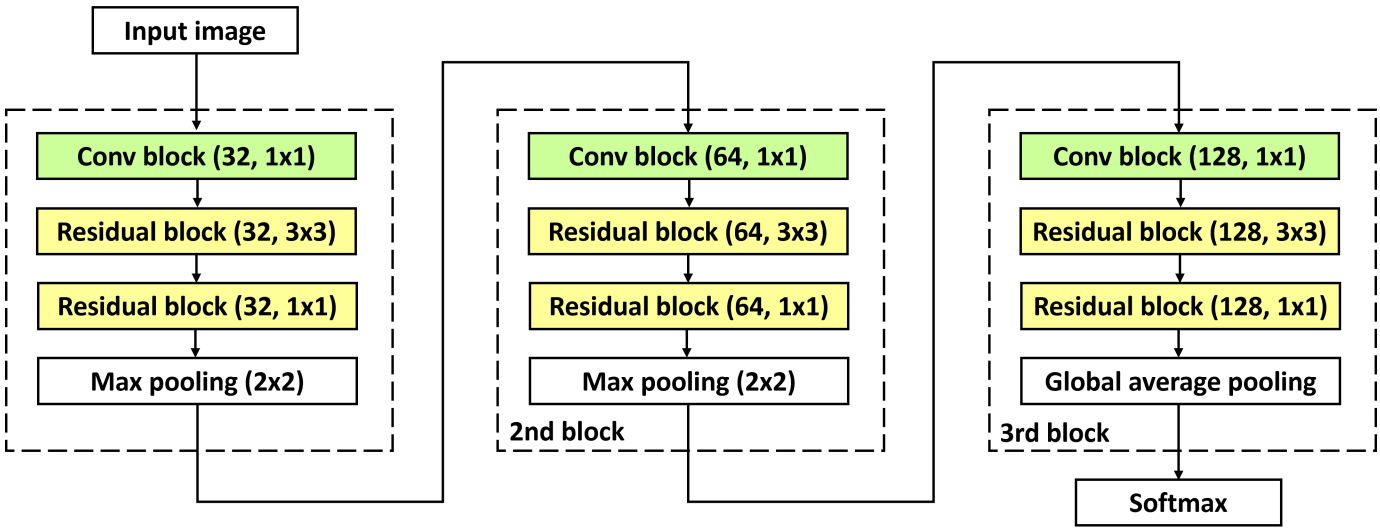
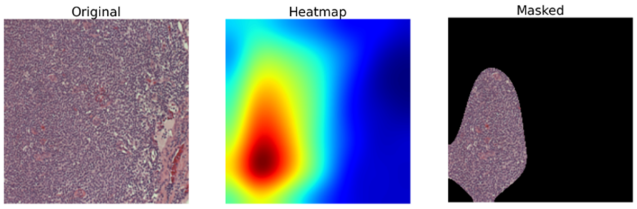


Fig. 6: Architecture of proposed ResNet.

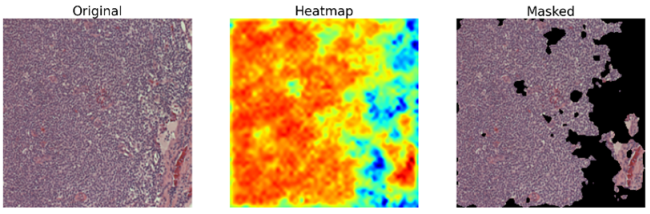
in the classification process doesn't work on our histological slides. Moreover, we were able to produce visual explanations for decisions.

We claim that explainable deep learning could be the meeting point between machine learning and physicians, who often refuse to use artificial intelligence techniques to treat their patients because of a lack of transparency in the process. For this reason, reassuring doctors about how machines come to a decision may be the only way to show impact of deep learning models in medical domain.

This project has some limitations. First, it is based on a dataset including only three lymphoma subtypes and use a dataset including more could be suitable. In addition, this dataset does not include for example molecular data, which has been demonstrated to be relevant to identify lymphoma subtypes. Second, there is no clinical confirmation about the reliability of the areas of interest indicated by the heatmaps. Further development of this work could be asking human hematopathologist to annotate regions of interest in images.



(a) For ResNet50-based architecture.



(b) For the proposed ResNet.

Fig. 7: Examples of heatmaps, related masks, and masked images.

REFERENCES

[1] Biaosheng Sheng, Mei Zhou, Menghan Hu, Qingli Li, Li Sun and Ying Wen, "A blood cell dataset for lymphoma classification using faster R-CNN.", Biotechnology and Biotechnological Equipment, 2020.

[2] Orlov NV, Chen WW, Eckley DM, Macura TJ, Shamir L, Jaffe ES, Goldberg IG., "Automatic classification of lymphoma images with transform-based global features.", IEEE Trans Inf Technol Biomed., 2010.

[3] Khoury JD et al., "The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms.", Leukemia, 2022.

[4] Janowczyk A, Madabhushi A., "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases.", J Pathol Inform, 2016.

[5] Rucha Tambe, Sarang Mahajan, Urmil Shah, Mohit Agrawal, and Bhushan Garware, "Towards Designing an Automated Classification of Lymphoma subtypes using Deep Neural Networks.", Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CODS-COMAD '19), 2019.

[6] Zhang X, Zhang K, Jiang M, Yang L., "Research on the classification of lymphoma pathological images based on deep residual neural network.", Technol Health Care., 2021.

[7] Syrykh, C., Abreu, A., Amara, N. et al., "Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning.", npj Digit. Med.,2020.

[8] Ammar Ammar et al., "Deep Learning for Lymphoma Detection on Microscopic Images.", Proceedings of the 4th International Conference on Life Sciences and Biotechnology (ICOLIB 2021), 2022.

[9] Guan Qingji, Huang Yaping, Zhong Zhun, Zheng Zhedong, Zheng Liang, Yang, Yi, "Thorax Disease Classification with Attention Guided Convolutional Neural Network.", ArXiv, 2019.

[10] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis." 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.", 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[12] He, K., Zhang, X., Ren, S., Sun, J., "Identity Mappings in Deep Residual Networks.", ArXiv, 2016.

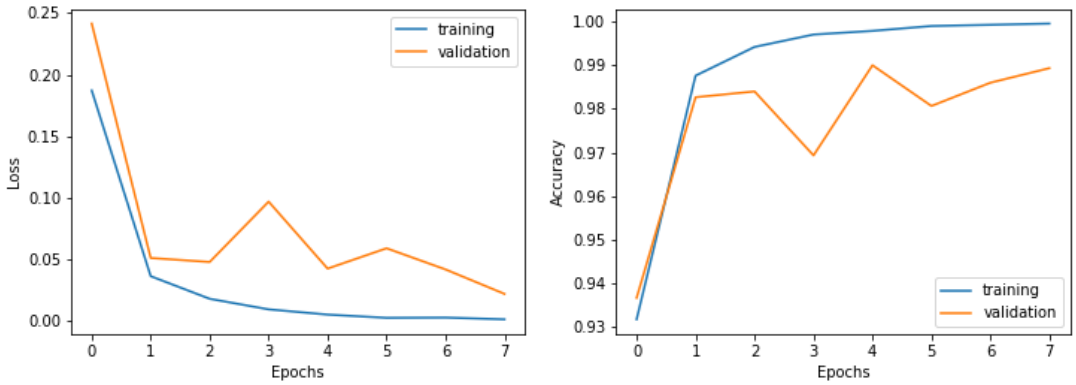


Fig. 8: Training and validation loss and accuracy versus number of epochs using ResNet50 on the dataset pre-processed to remove artifacts and augmented by rotating and splitting images.

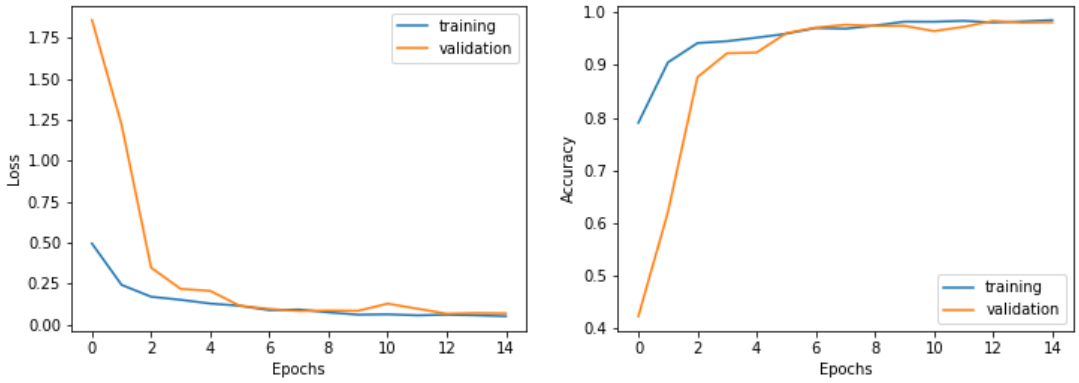


Fig. 9: Training and validation loss and accuracy versus number of epochs using the proposed ResNet on the dataset pre-processed to remove artifacts and augmented by dividing images into non-overlapping patches.

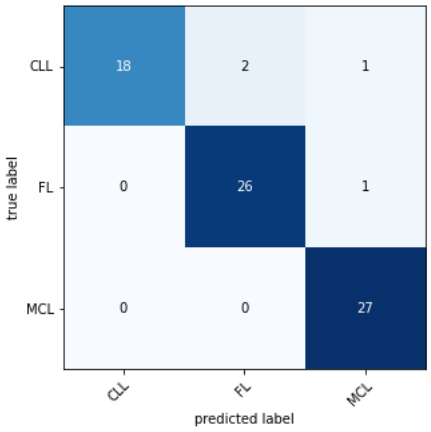
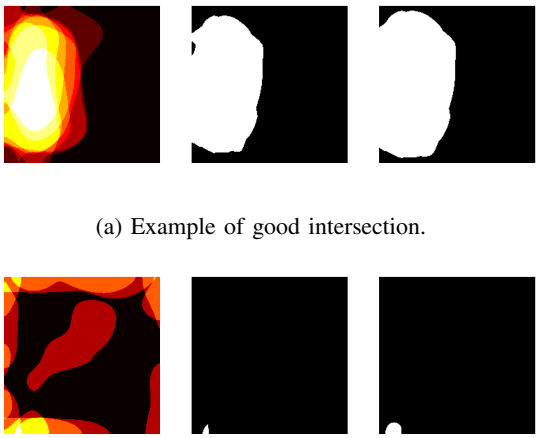


Fig. 10: Confusion matrix on test set for the proposed ResNet.



(a) Example of good intersection.

(b) Example of bad intersection.

Fig. 11: Intersection of masks obtatined from heatmaps.