Reliable Leukemia diagnosis and localization through Explainable Deep Learning

Marcello Di Giammarco[†], Benedetta Dukic[*], Fabio Martinelli[†],
Mario Cesarelli[‡], Fabrizio Ravelli[*], Antonella Santone[*], Francesco Mercaldo[*†]
[*]Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, Campobasso, Italy
{b.dukic, f.ravelli}@studenti.unimol.it, {francesco.mercaldo, antonella.santone}@unimol.it
[†]Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy
{fabio.martinelli, francesco.mercaldo}@iit.cnr.it
[‡]Department of Engineering, University of Sannio, Benevento, Italy
mcesarelli@unisannio.it

***Abstract*—Acute lymphoblastic leukemia is a cancer of the blood and bone marrow, it is the most common form of childhood cancer. In the United States, approximately 75% of people under age 20 diagnosed with leukemia are diagnosed with acute lymphoblastic leukemia. An estimated 400 people ages 15 to 19 in the US are diagnosed with the disease each year. In this paper we propose a Deep Learning network-based approach to detect acute lymphoblastic leukemia from images of blood cells. Additionally, the suggested approach can offer predictability through class activation mapping, which aims to automatically highlight the relevant and suspicious patterns in the image. We consider a method that uses the output of two separate class activation mapping techniques to determine whether the acute lymphoblastic leukemia prediction and localization can be regarded as resilient. Using a dataset of 6.099 blood cell images, we assess the efficacy of the suggested method and achieve an accuracy of 94%, demonstrating the usefulness of the proposed network for acute lymphoblastic leukemia detection and localization. Our method introduces also a similarity index aimed to "quantify" qualitative results coming from the heatmaps, in such a way as to improve the trustworthiness and reliability of the Artificial Intelligence for the medical staff.**

***Index Terms*—Machine Learning, Artificial Intelligence, Acute Lymphoblastic Leukemia, Visual Explainability**

I. Introduction

Acute lymphoblastic leukemia (ALL) is a malignant transformation and proliferation of white blood cells called lymphocytes. The hallmark of ALL involves chromosomal abnormalities and genetic alterations associated with the differentiation and proliferation of malignant cells. In 2019, it is estimated that approximately 6,000 new cases of ALL occurred in the US, which is less than 4% of all blood cancers. Roughly 50% of these ALL cases occur in children and represent greater than 30% of all pediatric cancers. The incidence of ALL follows a bimodal distribution, with the first peak occurring in childhood and the second peak occurring around the age of 50.

In children, survival rates for pediatric ALL have dramatically improved in the past 50 years, with cure rates exceeding 85% in children. However, resistance to therapy is common and for some children or adolescents, many therapies simply do not work, resulting in the use of multiple rounds of alternative chemotherapy. Beyond this, the current therapies often lead to long-term side effects (i.e. central nervous system impairment, cardiovascular issues, bone growth defects). Therefore, there is an urgent need to develop safer, more effective treatments for ALL, particularly for children who are refractory or resistant. It is also of fundamental importance to be able to identify the symptoms of leukemia as early as possible, in order to diagnose the disease and start the relative therapy. ALL is usually suspected when a test finds abnormal blood counts and leukemic cells, or blasts, appear in the blood. Then, the diagnosis is established by examination of the bone marrow via bone marrow aspiration and biopsy. ALL is diagnosed when the bone marrow aspirate and biopsy contains 20 percent or more immature cells called blasts, determined to be lymphoid in nature.

It is generally difficult to be certain of an ALL diagnosis simply by the appearance of cells under the microscope. Therefore, additional time-consuming laboratory tests are normally needed.

One important test is immunophenotyping (also called flow cytometry), which determines whether the cells are lymphoid (ALL) rather than myeloid (AML), based on proteins expressed in the leukemia cells. Immunophenotyping also determines whether they are T or B lymphocytes. In addition, chromosome testing, called cytogenetics, is a critical part of the evaluation that helps determine the appropriate course of treatment[1].

Recently, we have assisted a growing interest from both the industrial and academic world in the adoption of Deep Learning (DL) to face bioengineering-related challenges, from Alzheimer's diagnosis directly from brain computerized tomography [3] to the red blood cells, white blood cells and platelets counting from microscopic images [16]. As a matter of fact, currently, almost every device intended for medical imaging has a more or less extended image and signal analysis and for this reason, it is possible to analyze these data by exploiting and thus integrating artificial intelligence in real-world medical devices [6], [8], [9], [14], [15], [23].

[1]https://www.ucsfhealth.org/conditions/acute-lymphoblastic-leukemia/

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

For these reasons, in this paper, we propose a method aimed at detecting the presence of ALL from microscopic images related to blood cells. We resort to DL, by exploiting several convolutional neural networks (CNNs). Considering that one of the factors that prevent the introduction of artificial intelligence in the real world (especially in the medical one) is the lack of explainability, intending to understand the reasons behind a given prediction we resort to a class activation mapping technique, in particular the Gradient-weighted Class Activation Mapping (Grad-CAM) [22] and the Score-Weighted Visual Explanations Class Activation Mapping (Score-CAM) [24] to highlight the areas on the microscopic images that are symptomatic from the ALL disease from the DL model point of view[2]. Finally, this paper introduces the Structural Similarity Indeces Measure (SSIM) [3], a metric to measure the similarity between two given images. The novelty concerns the SSIM application i.e. between the previous two CAM algorithms; intending to "quantize" the qualitative features, guaranteeing better reliability and trustworthiness of the AI for the medics. The paper proceeds as follows: in the next section, we present the proposed method, in Section III the experimental analysis results are presented and discussed, the state-of-the-art literature is illustrated in Section IV and, finally, in the last section conclusions and future research lines are drawn.

II. The Method

In this section, the methodology applied for the detection of ALL from cytological images of blood cells is presented.
The main steps are shown in figure 1.
The first step, and one of the fundamental steps, is to obtain a dataset that is as representative as possible of the problem under investigation. In addition, concerning the medical context, the dataset must also be certified and validated to make the results obtained usable.
The next step consists in pre-processing the data contained in the dataset, to extrapolate or improve the features contained therein and eliminate any sources of noise and disturbance to the classification process. In the following, the architectures to be used in the training and testing phase are defined, together with the hyperparameters to be used, to select the best network-hyperparameter combination, and also to make a comparison between the different networks. The quantitative performance of the models is evaluated using the metrics of accuracy, precision, recall, AUC, F-measure and loss.
Finally, a qualitative analysis is carried out to provide explanations for the models obtained. This is performed using CAM algorithms, in particular Grad-CAM and Score-CAM, and SSIM.

A. Dataset and preprocessing

The dataset used in this paper was obtained from the Kaggle website, loaded by Larxel [4]. The dataset contains a total of 15135 images from 118 different subjects and is divided into two classes: Healthy and ALL.
The dataset was originally distributed for the proposed challenge at ISBI 2019. [4]
The task proposed in this challenge was to identify immature leukemic blasts from healthy cells, a difficult task as the cells appear morphologically similar.
Figure 2 shows representative images of the classes in the dataset.

The images contained in the dataset were processed appropriately before being used in the training and testing phase. First, the images were converted to 'png' format and resized to 256x256 pixels.
Subsequently, a centred zoom was performed on them, using a zoom factor of 0.6, to eliminate as much as possible of the background region and thus areas irrelevant to the classification process.
The reason for using the zoom instead of identifying the cell boundaries and subsequently cropping the image is that the cells under examination have extremely inhomogeneous contours and therefore this technique could have been counterproductive. Furthermore, this technique allows us to preserve information regarding the shape and size of the cells, which instead was compromised using cropping techniques.

B. Visual Explainability

CAM (Class Activation Mapping) algorithms, in particular Grad-CAM and Score-CAM, and structural similarity indices (SSIM) were used to provide explainability for the models obtained.

CAM-based algorithms make it possible to highlight the portions of the image that have contributed most to the classification, allowing us to understand the feature of the problem that is most discriminating, as well as highlighting possible regions not yet taken into account in current studies and thus guiding future developments. Moreover, the heatmaps are based on the VIRIDIS coloration [5], regarding the pattern importance during the classification process. The Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm exploits the back-propagation of individual class weights to highlight areas of interest. [22]
The Score-CAM (Score-weighted Class Activation Mapping) algorithm solves the gradient saturation and false-confidence problems of Grad-CAM using a different approach. [24]
Score-CAM uses a parameter defined as Channel-wise Increase of Confidence (CIC) to assess the contribution of each feature map based on the class score.

To further increase the level of explainability of the models, we compute the SSIM index, introduced in [13]. SSIM is an index that takes into account variations in contrast, brightness, and possible distortions relative to two versions of the same image. From this, the Model Robustness Structural Similarity Index (MR-SSIM) is derived. This index indicates how different two heatmaps produced by different CAM algorithms,

[2]https://www.lls.org/research/acute-lymphoblastic-leukemia-all
[3]https://www.imatest.com/docs/ssim/
[4]https://www.kaggle.com/andrewmvd/datasets
[5]https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

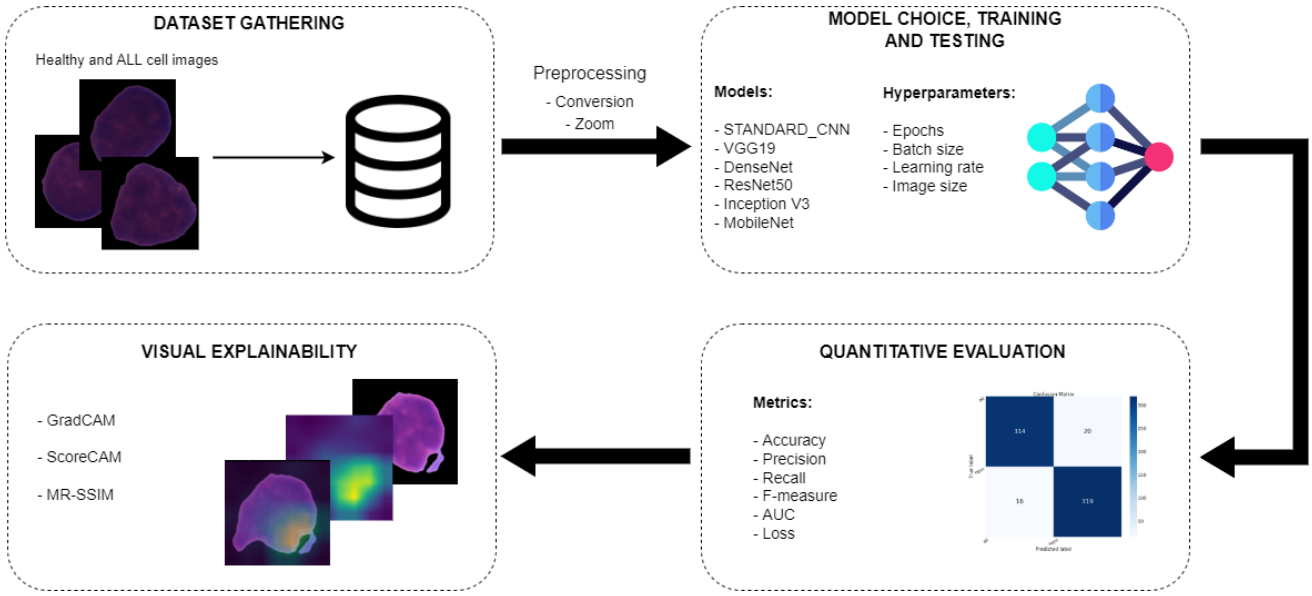

Fig. 1: Main steps of the proposed method

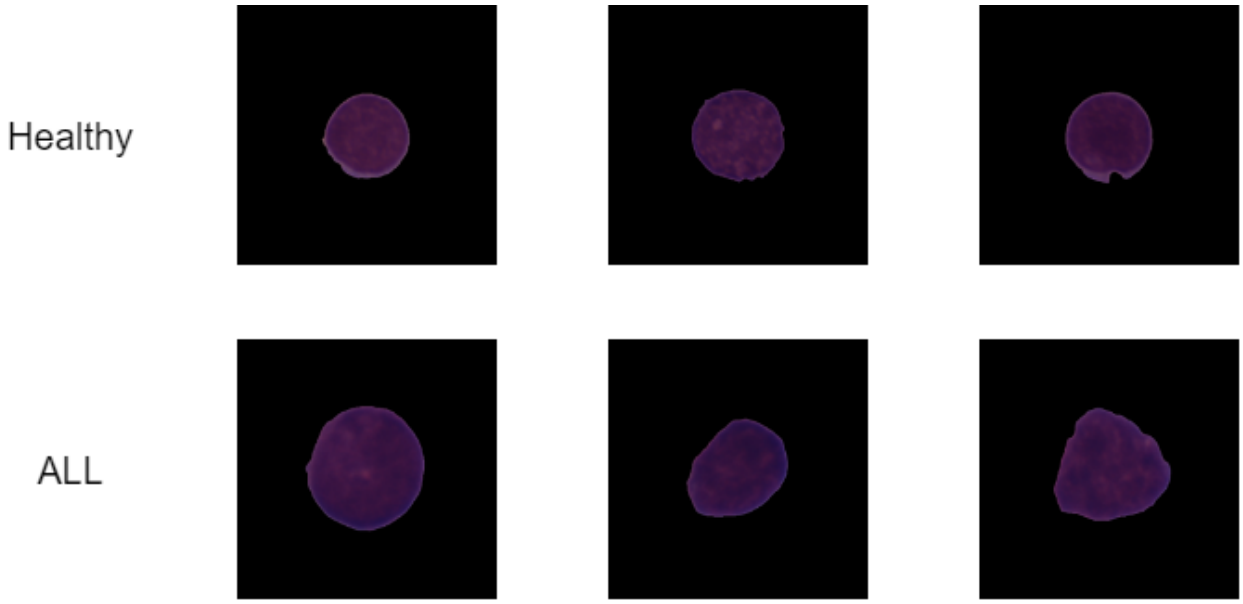

Fig. 2: Representative images of the dataset

using the same model, are. This index takes values between +1 and -1, where +1 indicates that the two images are equal. High values of MR-SSIM make it reasonable to assume that the highlighted portions of the images are indeed of interest as they are highlighted by different visual explanation algorithms.

III. Experimental Analysis and Results

In this section, we report the dataset under analysis and all the results, quantitative and qualitative, extracted from the images. In the final part, we discuss these results.

The used dataset derives from the Kaggle site, at the following URL[6]. This dataset is composed of three directories (training, testing and validation data), however, only the training_data directory is labeled to distinguish healthy subjects (hem) from the disease patients (all). Moreover, this directory is divided into three folders (fold_o, fold_1, and fold_2) which themselves contain the two classes of patients. These classes for each folder are unbalanced with a huge amount of healthy subjects, so we decided to balance the classes in terms of the sample number for a total of 6.099 images, and applying the 80-10-10 splitting we obtain the division of samples as follows:

[6]https://www.kaggle.com/datasets/andrewmvd/leukemia-classification

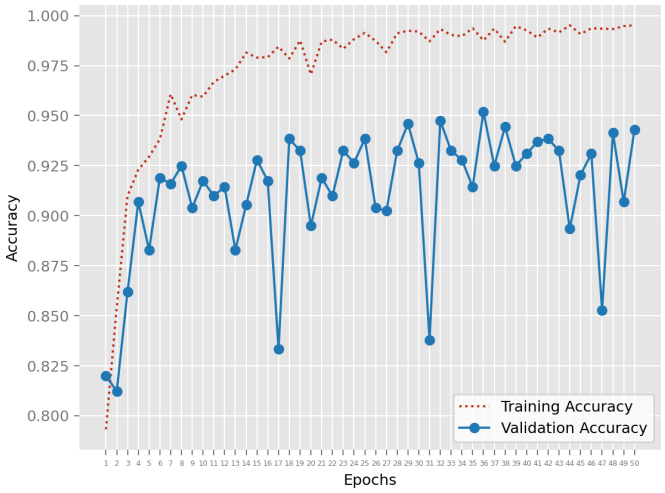

Fig. 3: Epoch-accuracy trend for the MobileNet model

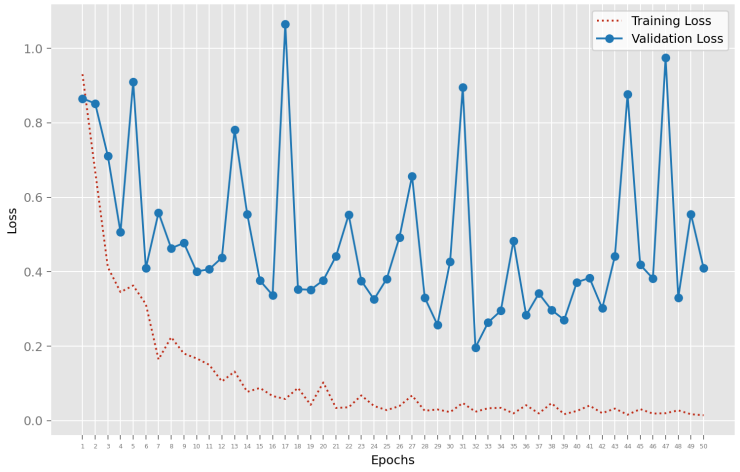

Fig. 4: Epoch-loss trend for the MobileNet model

- 80% of images (4.764) to the training dataset
- 10% of images (666) to the validation dataset.
- 10% of images (669) to the testing dataset.

After the application of the pre-processing technique cited in Section II, several CNNs are trained and tested on the dataset. The networks are: ResNet50 [21], DenseNet [28], VGG19 [25], Standard_CNN [2], InceptionV3 [26] and MobileNet [7]. The hyper-parameters settings are 50 epochs, 8 batch size, 0.0001 learning rate, and 224x224x3 image size. This optimal combination is found by testing different combinations on all networks. All training and tests were performed in a working environment with an MSI Intel Core i7 with 16 GB RAM.

In table I are reported the metrics of the networks in terms of accuracy, precision, recall, F-Measure, AUC, and loss.

As shown in Table I, the best performances are related to the MobileNet and DenseNet networks with 94% in accuracy precision and recall. Also, Inception V3, Standard_CNN and resNet 50 show good results, 92%, 88% and 82%respectively. From these results, it is possible to deduce that these networks (i.e. MobileNet and DenseNet) can correctly classify healthy and diseased blood cells. Vice versa, VGG19 reports the worse metrics, so are not able to distinguish the two image classes.

In Figures 3 and 4 are shown for MobileNet network the epoch-accuracy and epoch-loss trends, respectively.

In Figure 3 we can note good results during the training phase, and a slight decrease for the validation phase. The accuracy plot shows training accuracy (red dotted line) increases and plateaus, while validation accuracy (blue line) increases and stabilizes or decreases if overfitting occurs. This is an expected behavior and both training and validation accuracy confirm that the MobileNet model was able to learn the differences between images belonging to different classes. The opposite behavior is obtained from the (training and validation) loss, shown in Figure 4 and this is another confirmation that the model is correctly learning the differences between healthy cells and ALL cells. Small oscillations in the validation

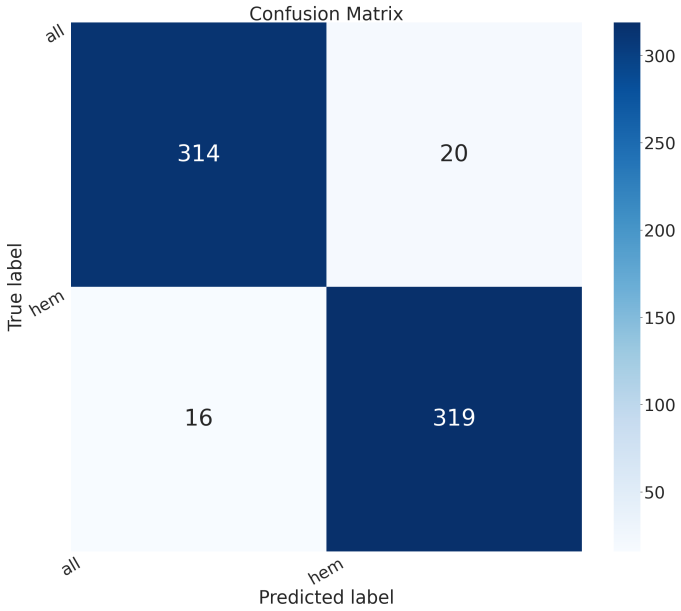

Fig. 5: Confusion matrix of the MobileNet model.

accuracy and loss during the training of a CNN are normal and can be attributed to various factors related to the training process, data characteristics, and model behavior.

In Figure 5 is reported the confusion matrix for the MobileNet network. The matrix, in Figure 5 reveals the good performance of the model, with higher values on the first diagonal of the matrix, and this means that objects labeled in a certain class are correctly predicted in the correct class.

A. CAMs algorithms evaluation

In this subsection, qualitative results are evaluated. This evaluation is based on the CAM algorithms application, i.e. Grad-CAM and Score-CAM, as mentioned in Section II.

CNN	Accuracy	Precision	Recall	F-Measure	AUC.	Loss
ResNet 50	0.82	0.82	0.82	0.82	0.87	1.80
DenseNet	0.94	0.94	0.94	0.94	0.98	0.22
VGG19	0.49	0.49	0.49	0.49	0.5	0.69
Standard_CNN	0.88	0.88	0.88	0.88	0.92	0.80
Inception V3	0.92	0.92	0.92	0.92	0.97	0.20
MobileNet	0.94	0.94	0.94	0.94	0.97	0.26

TABLE I: Metrics evaluation for tested DL models.

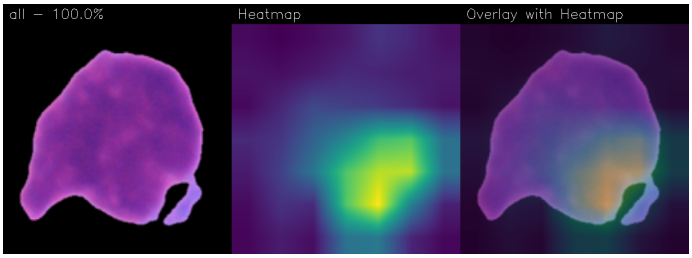

Fig. 6: Grad-CAM of the MobileNet model.

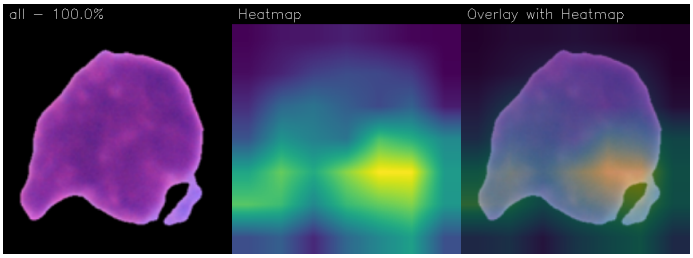

Fig. 8: Grad-CAM of the DenseNet model.

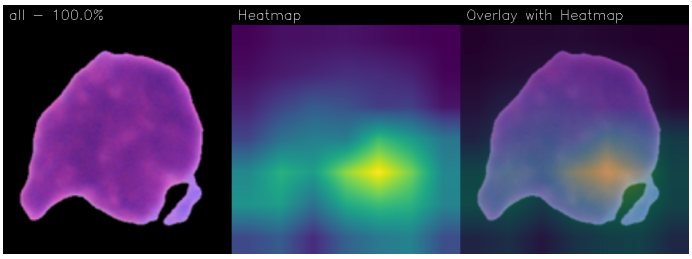

Fig. 7: Score-CAM of the MobileNet model.

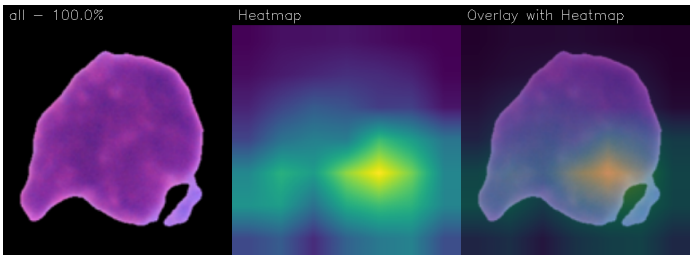

Fig. 9: Score-CAM of the DenseNet model.

Figures 6 and 7 are reported the Grad-CAM and Score-CAM generated by the MobileNet network, taking into account the same ALL image.

Starting from the left, the Figures are composed of the input images, the generated heatmaps, and the overlapping of the previous two.

Moreover, for the DenseNet network that reports optimal quantitative results, an evaluation regarding the qualitative point of view shows interesting heatmaps. Figures 8 and 9 are reported by the Grad-CAM and Score-CAM generated by the DenseNet network, respectively.

In general terms, all the overlapped heatmaps highlight the same ROI (Region Of Interest), improving the fact that different explainability and localization techniques (Grad-CAM and Score-CAM) with different models applications (MobileNet and DenseNet) reveal almost the same pattern, i.e. the lower right region. Moreover, the similarities regard not only the highlighted areas but also the heatmap coloration. Heatmaps generated by both methods use a color gradient (e.g., from blue to yellow/red) to indicate the importance of different regions. Red or warmer colors typically represent areas of high importance (ROIs), while blue or cooler colors represent areas of low importance. The gradient of colors used to indicate importance will be applied similarly, resulting in comparable visual patterns. This region presents a protuberance on the cell border, that can be related to the infant ALL presence. These kinds of protuberance and irregularity are linked with chromosomal abnormalities and genetic alteration due to the infant ALL. So, these similar heatmaps increase the reliability and trustworthiness of AI for medics; passing from a "black box" model to a visually explainable approach, with not only a quantitative classification, but also mainly a localization of the relevant pattern for the decision-making, contributing to a correct diagnosis.

In addition, the authors try to quantify the qualitative results and improve the model robustness, by introducing the MR-SSIM. Table II shows the average similarity value between the two sets of Grad-CAM and Score-CAM heatmaps for each class and models.

Table II compares the heatmaps that are activated with the Grad-CAM and Score-CAM algorithms on the same model (i.e., the MobileNet and DenseNet). The MR-SSIM indices

Similarity Index for heatmaps sets		
Models	**Classes**	**MR-SSIM**
MobileNet	hem	0.888
	all	0.892
DenseNet	hem	0.948
	all	0.945

TABLE II: MR-SSIM for the heatmaps sets..

report 0.88 for MobileNet and 0.94 for DenseNet; this means that the generated heatmaps coming from two different CAMs are very similar and identify the same areas with slight differences. From the medical point of view, these analyses represent a strong method to identify and locate ALL in infants, with two robust and explainable deep-learning models to support it for a correct diagnosis.

The combination of the CAMs and the MR-SSIM offers benefits such as improved interpretability by highlighting crucial areas within images, enhanced diagnostic accuracy through precise region localization, and better image quality assessment by measuring structural similarities to reference images, ultimately leading to more reliable and effective medical diagnoses; and represent the main novelty of our work.

IV. Related Work

In this section, we provide an overview of the existing literature that pertains to the utilization of DL in the context of ALL detection. Following this, we delve into a detailed discussion of these research papers.

The dataset used by Rehman et al. [19] is categorized into three distinct subtypes of ALL, namely L1, L2, and L3, along with a healthy category named normal and it is established through collaboration with hematologists. The innovative aspect of this approach is that it accepts bone marrow images as input, carries out segmentation, and distinguishes between normal and affected marrow or its subtypes reaching an accuracy of 0.9778. The architecture proposed in this article, which is implemented in MATLAB with computer vision toolbox and Alexnet model on GPU, also provides the segmentation of different blasts.

In the study developed by Zakir Ullah et al. [27], the training process is initiated from the ground up. The use of a network that is not pre-trained allows it to consider and extract features that could be different from the ones extracted from the pre-trained network. So This process aims to determine parameter values that are highly pertinent and result in improved convergence for the network. Different data augmentation techniques are used with the aim of avoiding overfitting and balancing the dataset. This work also consists of the ECA module that extracts information from each channel of VGG16 results and then combines them using a weighted sum. This enables the DL model to assign greater importance to particular elements within the input images. When comparing the performance of the VGG16 model with and without the ECA module, it became evident that the attention mechanism significantly enhances model accuracy, exhibiting a mean accuracy of 0.911.

In the article of Pansombut T et al. [**?**], the researchers present a CNN named ConVNet for image-based classification that employs traditional machine learning methods like Support Vector Machines, Multi-Layer Perceptron, and Random Forest for feature-based classification. Various features are extracted from cell images to aid in classification. The performance of ConVNet is described by an average accuracy of 0.8174.

The study [10] introduces the ViT-CNN which is based on two different methods used to extract and combine features from cell images. A vision transformer model and CNN are fused together in order to diagnose ALL. The ViT-CNN is able to reach a classification accuracy of 99.03%

In [5], authors introduce a complete database of 15114 total images made up in the All India Institute of Medical Sciences (AIIMS) of New Delhi. It consists of images of cancer cells collected from patients who suffered from B-ALL and others from healty ones as controll for a total of 10661 cell images as training set, 1867 as preliminary test set and 2586 as final test set. This rich and accurated dataset was utilised for performing a medical imaging challenge in IEEE International Symposium on Biomedical Imaging (ISBI)-2019. In this occasion, many teams brought in works based on CNN architectures and carry out a score called F_1. The highest score, achieved by Yongsheng Pan et al. [18] is due to a neighborhood-correction algorithm as a solution to this challenge. It involves three main steps: the first is the production of initial labels and feature maps for test data with a fine tuned pre-trained ResNet; the second is the construction of a Fisher vector for every single image using the previous found feature maps and the final one consists of adjusting the initial labels evaluating similarity with neighbors. This method reached 0.910 of the weighted F_1 which is a balanced metric in order to take in count the imbalance of the dataset.

The result obtained by Liu Y. et al. [12] is a F_1 of 0.876. In this study, a strategy based on two-stage training is adopted. Firstly two Inception ResNets are used to be trained on a subdivided dataset (A and B) and then the resulting models from each one are matched and fine-tuned utilizing the complete dataset.

This dataset is also used by other researchers as in [20] where the previous one is subjected to data augmentation to remove the imbalance due to the distribution of cancer and healthy cells. So they shape a dataset made up of 12000 cells proportionally divided in ALL and HEM. They get a resulting F_1 of 95.43% through a customized CNN named ALL-NET which consists of a max pooling next to every convolution layer, the batch normalization to ensure that the data in motion are normalized, and a final dropout to keep away from overfitting.

In [1], the Resnet101 ensemble model is a combination of several Resnet-101 models that are trained to identify ALL in microscopic images through a majority voting strategy. The most effective set of algorithm hyperparameters for the pre-trained Resnet-101 model is identified using the Taguchi experimental method. This method allows for a systematic and efficient exploration of different combinations of hyperparameter values to optimize the performance of the Resnet-101 model. They achieve an accuracy of 85.11% and an F_1-score of 88.94%.

In the work introduced by Mondal et al. [17] the initial dataset from C_NMC_2019 is transformed into a balanced one through various steps of pre-processing methods and the original image size of 450 x 450 is reduced to 300 x 300 in order to emphasize its central part with the result of reaching a better ability to discriminate important features. Proposing a weighted ensemble model, they achieve an F_1-score of 89.7%. The use of fine-tuning on various pre-trained CNN models allow authors to build up a weighted ensemble model to improve the classification. It is also reported a qualitative evaluation of results based on the visualization of Grad-CAM.

In [11] Lamberti uses a Random Forest method to classify images coming from two different datasets. The attention is concentrated on the possible features that can be extracted during the analysis such as color, shape and texture. In order to make the process suitable for every image regardless of the original dataset they belong, pre-processing is needed. Results consist of an accuracy of 100% in the first dataset and a F_1-score of 90.1% for the second one.

Table III provides a comparison between the works discussed in this section and the approach presented in this paper. It highlights the key outcomes, the different datasets used, and if the explainability is considered by authors for what concerns automatic disease localization.

V. Conclusion and Future Work

In conclusion, this paper aims to provide an automated method for ALL detection in infants considering cytological images of the blood cell to evaluate this disease. From the results, the applied CNNs respond with good quantitative results (94% in accuracy, precision and recall), in particular MobileNet and DenseNet architecture. This work evaluates mainly the qualitative aspect of tested models, introducing CAM algorithms to visual localization of features in the images that are related to the network classification and the abnormal formation over the cells. Moreover, the application of two distinguished CAMs increases the trustworthiness and reliability of AI in healthcare. This means that this technology doesn't substitute human decisions, but it helps medics during the diagnosis-making process, guaranteeing a practical implementation.

Future research plans to focus on other blood and Lymphatic disorders, focusing on the visual explainability to transform the AI into Explainable-AI (XAI), including applications of several CAMs and the similarity index. Parallel to this, another field of research includes the use of algorithms aimed at increasing the security of networks by applying various techniques such as GANs (Generative Adversarial Networks).

Acknowledgment

This work has been partially supported by EU DUCA, EU CyberSecPro, SYNAPSE, PTR 22-24 P2.01 (Cybersecurity) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU projects, by MUR - REASONING: foRmal mEthods for computAtional analySis for diagnOsis and progNosis in imagING - PRIN, e-DAI (Digital ecosystem for integrated analysis of heterogeneous health data related to high-impact diseases: innovative model of care and research), Health Operational Plan, FSC 2014-2020, PRIN-MUR-Ministry of Health, the National Plan for NRRP Complementary Investments D^3 4 Health: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care, Progetto MolisCTe, Ministero delle Imprese e del Made in Italy, Italy, CUP: D33B22000060001 and FORESEEN: FORmal mEthodS for attack dEtEction in autonomous driviNg systems CUP N.P2022WYAEW.

References

[1] Chen, Y.M., Chou, F.I., Ho, W.H., Tsai, J.T.: Classifying microscopic images as acute lymphoblastic leukemia by resnet ensemble model and taguchi method. BMC Bioinformatics **22** (01 2022). https://doi.org/10.1186/s12859-022-04558-5

[2] Di Giammarco, M., Iadarola, G., Martinelli, F., Mercaldo, F., Ravelli, F., Santone, A.: Explainable deep learning for alzheimer disease classification and localisation. In: International Conference on Applied Intelligence and Informatics, 1-3 September, Reggio Calabria, Italy. in press (2022)

[3] Di Giammarco, M., Iadarola, G., Martinelli, F., Mercaldo, F., Ravelli, F., Santone, A.: Explainable deep learning for alzheimer disease classification and localisation. In: Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings. pp. 129–143. Springer (2023)

[4] Gupta, A., Gupta, R.: Isbi 2019 c-nmc challenge: Classification in cancer cell imaging. Select Proceedings (2019)

[5] Gupta, R., Gehlot, S., Gupta, A.: C-nmc: B-lineage acute lymphoblastic leukaemia: A blood cancer dataset. Medical Engineering & Physics **103**, 103793 (03 2022). https://doi.org/10.1016/j.medengphy.2022.103793

[6] He, H., Yang, H., Mercaldo, F., Santone, A., Huang, P.: Isolation forest-voting fusion-multioutput: A stroke risk classification method based on the multidimensional output of abnormal sample detection. Computer Methods and Programs in Biomedicine p. 108255 (2024)

[7] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[8] Huang, P., Li, C., He, P., Xiao, H., Ping, Y., Feng, P., Tian, S., Chen, H., Mercaldo, F., Santone, A., et al.: Mamlformer: Priori-experience guiding transformer network via manifold adversarial multi-modal learning for laryngeal histopathological grading. Information Fusion **108**, 102333 (2024)

[9] Huang, P., Xiao, H., He, P., Li, C., Guo, X., Tian, S., Feng, P., Chen, H., Sun, Y., Mercaldo, F., et al.: La-vit: A network with transformers constrained by learned-parameter-free attention for interpretable grading in a new laryngeal histopathology image dataset. IEEE Journal of Biomedical and Health Informatics (2024)

[10] Jiang, Z., Dong, Z., Wang, L., Jiang, W.: Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model. Computational Intelligence and Neuroscience **2021** (2021)

[11] Lamberti, W.F.: Classification of white blood cell leukemia with low number of interpretable and explainable features. arXiv preprint arXiv:2201.11864 (2022)

Authors	Dataset	Methods	Localization	Results
Rehman et al. (2018)	330 images from Amreek Clinical Laboratory Saidu Sharif Swat KP Pakistan	Naïve Baysian KNN SVM Proposed CNN	No	0.783 0.804 0.909 0.977
Jiang et al. (2021)	40000 images from C_NMC_2019 Dataset	Vision transformer EfficentNet ViT-CNN	No	0.989 0.951 0.990
Zakir Ullah et al. (2021)	26802 images from C_NMC_2019 Dataset	VGG16 + Attention VGG16	Yes	0.911 0.855
Pansombut T et al. (2019)	2420 images from All_IDB and ASH Image Bank	ConVNet SVM-GA MPL Random Forest	No	0.817 0.816 0.761 0.784
Pan et al. (2019)	10661 images from C_NMC_2019 Dataset	Neighborhood-correction algorithm with fine-tuned ResNets	No	0.910 (Weighted)
Liu et al. (2019)	10661 images from C_NMC_2019 Dataset	Fine-tuning of Inception Fine-tuning of ResNets	No	0.876 (Weighted)
Sampathila et al. (2022)	12000 images from C_NMC_2019Dataset	ALL-NET	No	0.954 (Weighted)
Chen et al. (2021)	12528 images from C_NMC_2019 Dataset	Resnet 101-9	No	0.851 0.889(Weighted)
Mondal et al. (2021)	15114 images from C_NMC_2019 Dataset	VGG-16 Xception MobileNet InceptionResNet-V2 DenseNet-121 WEN	Yes	0.851 0.859 0.843 0.837 0.829 0.898
Lamberti (2022)	390 images from All_IDB2 10661 images from C_NMC_2019 Dataset	Random Forest	No	0.100 0.901 (Weighted)
Our method (2024)	6009 images from C_NMC_2019 Dataset	ResNet 50 DenseNet VVG19 Standard_CNN Inception V3 MobileNet	Yes	0.820 0.940 0.490 0.880 0.920 0.940

TABLE III: Comparison between the proposed method and the current leading methods found in the state of the art

[12] Liu, Y., Long, F.: Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning. In: ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings, pp. 113–121. Springer (2019)

[13] Mercaldo, F., Di Giammarco, M., Apicella, A., Di Iadarola, G., Cesarelli, M., Martinelli, F., Santone, A.: Diabetic retinopathy detection and diagnosis by means of robust and explainable convolutional neural networks. Neural Computing and Applications pp. 1–13 (2023)

[14] Mercaldo, F., Di Giammarco, M., Ravelli, F., Martinelli, F., Santone, A., Cesarelli, M.: Triad: A deep ensemble network for alzheimer classification and localisation. IEEE Access (2023)

[15] Mercaldo, F., Di Giammarco, M., Ravelli, F., Martinelli, F., Santone, A., Cesarelli, M.: Alzheimer's disease evaluation through visual explainability by means of convolutional neural networks. International Journal of Neural Systems **34**(2), 2450007–2450007 (2024)

[16] Mercaldo, F., Martinelli, F., Santone, A., Cesarelli, M.: Blood cells counting and localisation through deep learning object detection. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4400–4409. IEEE (2022)

[17] Mondal, C., Hasan, M.K., Ahmad, M., Awal, M.A., Jawad, M.T., Dutta, A., Islam, M.R., Moni, M.A.: Ensemble of convolutional neural networks to diagnose acute lymphoblastic leukemia from microscopic images. Informatics in Medicine Unlocked **27**, 100794 (2021). https://doi.org/https://doi.org/10.1016/j.imu.2021.100794, https://www.sciencedirect.com/science/article/pii/S235291482100263X

[18] Pan, Y., Liu, M., Xia, Y., Shen, D.: Neighborhood-correction algorithm for classification of normal and malignant cells. In: ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings. pp. 73–82. Springer (2019)

[19] Rehman, A., Abbas, N., Saba, T., Rahman, S.I.u., Mehmood, Z., Kolivand, H.: Classification of acute lymphoblastic leukemia using deep learning. Microscopy Research and Technique **81**(11), 1310–1317 (2018)

[20] Sampathila, N., Chadaga, K., Goswami, N., Chadaga, R.P., Pandya, M., Prabhu, S., Bairy, M.G., Katta, S.S., Bhat, D., Upadya, S.P.: Customized deep learning classifier for detection of acute lymphoblastic leukemia using blood smear images. Healthcare **10**(10) (2022). https://doi.org/10.3390/healthcare10101812, https://www.mdpi.com/2227-9032/10/10/1812

[21] Sankupellay, M., Konovalov, D.: Bird call recognition using deep convolutional neural network, resnet-50. In: Proc. Acoustics. vol. 7, pp. 1–8 (2018)

[22] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

[23] Sun, X., Li, L., Mercaldo, F., Yang, Y., Santone, A., Martinelli, F.: Automated intention mining with comparatively fine-tuning bert. In: Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval. pp. 157–162 (2021)

[24] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)

[25] Wen, L., Li, X., Li, X., Gao, L.: A new transfer learning based on vgg-19 network for fault diagnosis. In: 2019 IEEE 23rd international conference on computer supported cooperative work in design (CSCWD). pp. 205–209. IEEE (2019)

[26] Xia, X., Xu, C., Nan, B.: Inception-v3 for flower classification. In: 2017 2nd international conference on image, vision and computing (ICIVC). pp. 783–787. IEEE (2017)

[27] Zakir Ullah, M., Zheng, Y., Song, J., Aslam, S., Xu, C., Kiazolu, G.D., Wang, L.: An attention-based convolutional neural network for acute lymphoblastic leukemia classification. Applied Sciences **11**(22), 10662 (2021)

[28] Zhu, Y., Newsam, S.: Densenet for dense flow. In: 2017 IEEE international conference on image processing (ICIP). pp. 790–794. IEEE (2017)