Assessing the Impact of Data Augmentation and a Combination of CNNs on Leukemia Classification

Maila L. Claro[1], Rodrigo de M. S. Veras[1], Andre M. Santana[1], Luis Henrique S. Vogado[1], Geraldo Braz Junior[2], Fatima N. S. de Medeiros[3], Joao Manuel R. S. Tavares[4]

claromaila@ufpi.edu.br, rveras@ufpi.edu.br, andremacedo@ufpi.edu.br, lhvogado@ufpi.edu.br, geraldo@nca.ufma.br, fsombra@ufc.br and tavares@fe.up.pt

[a1]*Departamento de Computação da Universidade Federal do Piaui, Teresina, Brasil*
[2]*Departamento de Informatica da Universidade Federal do Maranhão, São Luis, Brasil*
[3]*Departamento de Engenharia de Teleinformatica, Universidade Federal do Ceara - Fortaleza, Brasil*
[4]*Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Portugal*

Abstract

An accurate early-stage leukemia diagnosis plays a critical role in treating and saving patients' lives. The two primary forms of leukemia are acute and chronic leukemia, which is subdivided into myeloid and lymphoid leukemia. Deep learning models have been increasingly used in computer-aided medical diagnosis (CAD) systems developed to detect leukemia. This article assesses the impact of widely applied techniques, mainly data augmentation and multilevel and ensemble configurations, in deep learning-based CAD systems. Our assessment included five scenarios: three binary classification problems and two multiclass classification problems. The evaluation was performed using 3,536 images from 18 datasets, and it was possible to conclude that data augmentation techniques improve the performance of convolutional neural networks (CNNs). Furthermore, there is an improvement in the classification results using a combination of CNNs. For the binary problems, the performance of the ensemble configuration was superior to that of the multilevel configuration. However, the results were statistically similar in multiclass scenarios. The results were promising, with accuracies of 94.73% and 94.59% obtained using multilevel and ensemble configurations in a scenario with four classes. The combination of methods helps to reduce the error or variance of

the predictions, which improves the accuracy of the used deep learning-based model.

Key words: Image Classification, Deep Learning, Ensemble, Leukemia, Multilevel

1. Introduction

Bone marrow occupies the bone cavity, where blood cells are produced. It contains the cells that give rise to red blood cells, known as erythrocytes, platelets, and white blood cells, also known as leukocytes (Souza and Gorini, 2006). The latter cells actively participate in the human immune system and help it to defend the body against invaders. Progenitor cells in the marrow, also known as stem cells or precursor cells, produce an average of 100 million leukocytes per day. These leukocytes help the body to combat and eliminate microorganisms and chemical structures that are strangers to it through their capture, i.e., phagocytosis or through the production of antibodies. One of the diseases that affect the functioning of the bone marrow is leukemia (Travlos, 2006).

Leukemia is a malignant disease of the white blood cells, usually of unknown origin. Its main characteristic is the accumulation of diseased cells in the bone marrow, which replace normal blood cells. A blood cell that has not yet reached maturity undergoes a genetic mutation that turns it into a cancer cell in leukemia. This abnormal cell does not operate properly, and it multiplies faster and has a shorter lifespan than of normal cells. Hence, the abnormal cancer cells replace healthy blood cells in the bone marrow.

The American Cancer Society (ACS) (Society, 2021) estimated that there would be 61,090 new cases of leukemia in 2021, with approximately 23,660 deaths; in particular, there would be 35,530 male cases and 25,560 female cases, leading to 13,900 male deaths and 9,760 female deaths.

The types of leukemia can be classified according to the worsening speed of the disease. Hence, the condition can be of the chronic type, which usually gets worse slowly, or of the acute type, which usually gets worse quickly. The types of leukemia can also be classified based on the kind of white blood cells they affect: lymphoid or myeloid cells (Mrozek et al., 2004). Thus, the main types of leukemia are acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic myeloid leukemia (CML), and chronic lymphocytic leukemia (CLL). Acute leukemia affects mainly children, and chronic

leukemia tends to affect adults and the elderly (Souza and Gorini, 2006).

Each type of leukemia has an appropriate treatment; therefore, a diagnosis in the early stage of the disease is demanded to provide the proper treatment successfully. On the other hand, the main treatments for more advanced disease phases aim to destroy the leukemic cells so that the bone marrow returns to produce normal cells. Figure 1 shows examples of the blood slide images used in the experiments of the current study, mainly ALL, AML, chronic leukemia, and healthy blood slides (HBS).

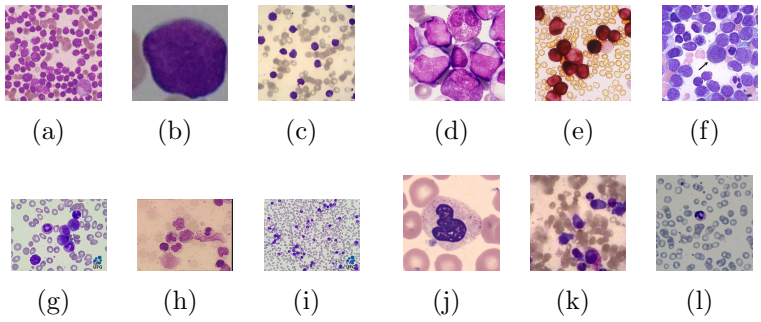


Figure 1: Examples of images used in this study: (a-c) ALL images, (d-f) AML images, (g-i) chronic leukemia images and (j-l) HBS images.

Deep learning models have been increasingly used in computer-aided medical diagnosis (CAD) systems (Puttagunta and Ravi, 2021). In particular, convolutional neural networks (CNNs) can learn hierarchical representations, from more general features in the first convolutional layers, to more semantic features in the last few layers. Currently, CNNs are one of the most effec-tive techniques used in medical imaging-based diagnosis (Upreti et al., 2021). Researchers have been seeking to increase the generalizability of CNNs, particularly based on techniques of data augmentation and the combination of CNNs in ensemble (Dietterich, 2000) and multilevel (Szegedy et al., 2015) configurations.

In this study, techniques that are widely used in CNN based CAD systems were evaluated, mainly data augmentation and ensemble and multilevel configurations (Kim et al., 2021). Therefore, seven CNNs were studied using different techniques of data augmentation and ensemble and multilevel configurations. According to five leukemia classification scenarios, the analysis was performed using 3,536 images from 18 heterogeneous datasets. Three of these scenarios are binary classification problems: leukemia vs. HBS, ALL

vs. HBS, and AML vs. HBS. The other two scenarios are multiclass classification problems: ALL vs. AML vs. HBS and ALL vs. AML vs. HBS vs. other types.

The main contributions of this article are the following: the identifications of the datasets that are widely used for leukemia classification, the introduction of five scenarios for the classification of different types of leukemia, the evaluation of the performance achieved by various CNN-based models on leukemia classification, the assessment of the impact of multiple data augmentation techniques on the classification performance, and assessment and comparison of the improvements achieved by multilevel and ensemble model configurations.

This article is organized as follows. Section 2 presents related work. Section 3 describes the used materials and methods, such as the used datasets, the employed techniques of data augmentation, the evaluated network architectures, the used ensemble and multilevel configurations, and, finally, the adopted evaluation metrics. Sections 4 and 5 present the achieved results and a comparison of them against the ones of previous works found in the literature, respectively. Finally, the conclusions and possibilities for future work are pointed out in Section 6.

2. Related work

This section presents studies that have been developed for leukemia detection. Taking into account the applied methodology, we identified traditional methods (Singhal and Singh, 2016; Madhukar et al., 2012; Goutam and Sailaja, 2015; Rawat et al., 2017; Laosai and Chamnongthai, 2018) and methods based on deep learning (Thanh et al., 2018; Vogado et al., 2018; Loey et al., 2020; Shafique and Tehsin, 2018). Traditional methods comprise several steps, such as image pre-processing, segmentation, feature extraction, and classification. On the other hand, procedures based on deep learning usually apply CNNs. This kind of procedures aims to design and build a CNN that takes the input image and returns the output class without going through the different challenging tasks involved in traditional methods, such as the segmentation of regions of interest (ROIs) and feature extraction.

We performed the current state-of-art review using the Scopus, Web of Science, and IEEE Xplore databases. The following search strings were used to search the engineering and computer science fields: “leukemia acute classification,” “white blood cell classification” and “blood smear leukemia clas-

sification". The first string returned 337 articles, the second string returned 271 articles, and the last string returned 59 articles. We then analyzed the titles and abstracts of all returned articles to eliminate the repeated documents, documents with non-automatic classification methods, and documents whose goal was just leukocyte segmentation as in (Jiang et al., 2003; Abbas and Mohamad, 2014; Arslan et al., 2014; Vogado et al., 2016; Li et al., 2016).

Given the accelerated evolution of CAD systems seen in the last decade, just the works published after 2012 were selected. About older approaches, the articles written by Mishra et al. (2016), Anilkumar et al. (2020) and Khan et al. (2020) can be suggested. Mainly, Mishra et al. (2016) performed a comparative analysis of methods published between 2002 and 2014. Anilkumar et al. (2020) presented a review of works published between 2004 and 2020 in the area of blood slide image processing, highlighting the automatic detection of leukemia. On the other hand, Khan et al. (2020) investigated works based on traditional machine learning and deep learning models. These authors collected 80 articles published in journals, books, conferences and online from 2014 to 2020.

The found articles can be classified based on two criteria: (1) the methodology applied and (2) the number of leukemia types present in the used image dataset. Due to the number of articles that were found, they were divided them into six groups of approaches that have been proposed to detect leukemia: (1) the approaches that differentiate between images with leukemia and healthy patterns, regardless of the type of leukemia; (2) the approaches that differentiate blood slides with ALL from healthy slides; (3) the approaches that differentiate images with AML from healthy images; (4) the approaches that differentiate images with CLL from healthy images; (5) the approaches developed to classify types of acute leukemia (ALL and AML) and healthy images; and finally, (6) the approaches that differentiate the four main types of leukemia (ALL, AML, CLL, and CML) from healthy images. Table 1 summarizes the main characteristics of the selected sate-of-art articles.

In the group of works performing binary classification, the input images are classified into healthy slides and images with leukemia, regardless of the type of leukemia. In the study performed by Thanh et al. (2018), a CNN-based method is proposed for the ALL-IDB1 dataset (Labati et al., 2011), which has 108 images: 59 healthy images and 49 images with ALL. These authors used techniques of data augmentation such as reflection, translation, rotation, and shear. The proposed method achieved 96.6% of ac-

Table 1: Summary of the identified state-of-the-art articles in terms of the used descriptor(s), classifier, number of datasets, number of images and obtained accuracy.

Work	Descriptor(s)	Classifier	Number of datasets	Images	Accuracy (%)
Leukemia - HBS					
Thanh et al. (2018)	Proposed CNN	CNN	1	108	96.60
Vogado et al. (2018)	CNNs VGG-f, AlexNet, CaffeNet	SVM	8	1,268	99.76
Loey et al. (2020)	CNN *AlexNet*	CNN	2	564	100
Vogado et al. (2021)	CNN *LeukNet*	CNN	18	3.536	98.61
ALL - HBS					
Singhal and Singh (2016)	Texture	SVM	1	260	93.80
Shafique and Tehsin (2018)	CNN AlexNet	CNN	2	368	99.50
Ahmed et al. (2019)	Proposed CNN	CNN	2	354	88.25
Pansombut et al. (2019)	CNN ConVNet	CNN	2	363	81.74
Gehlot et al. (2020)	CNN SDCT AuxNet$^{\theta}$	CNN	1	15,114	93.40
Das and Meher (2021)	MobileNetV2+ ResNet18	CNN	2	368	97.92 and 96.00
Khandekar et al. (2021)	Yolov4	Yolov4	2	1,108	mAP:96.06;98.70
Zakir Ullah et al. (2021)	VGG16-ECA	CNN	1	15,114	91.10
Karar et al. (2022)	CNN	GAN	1	368	98.65
Rodrigues et al. (2022)	ResNet-50V2-GA	CNN	1	260	98.46
Abhishek et al. (2022)	Vgg16	CNN	2	608	97.00
Rastogi et al. (2022)	LeuFeatx	ETC	1	260	96.15
AML - HBS					
Madhukar et al. (2012)	Texture	SVM	1	50	93.50
Goutam and Sailaja (2015)	Texture	SVM	1	90	98.00
Dasariraju et al. (2020)	Shape and Color	Random Forest	1	1,274	92.99
CML - HBS					
Khosla and Ramesh (2018)	Deep Features	CNN	1	67	97.60
ALL - AML - HBS					
Rawat et al. (2017)	Geometrical, color and texture	GA-SVM	1	240	99.50
Laosai and Chamnongthai (2018)	Shape, color distribution, texture and number of nucleoli	SVM	2	500	99.85
Tran et al. (2018)	CNN LeukemiaNet	CNN	2	141	97.20
Claro et al. (2020)	CNN Alert Net-RWD	CNN	16	2,415	97.18
Karar et al. (2022)	CNN	GAN	2	445	95.50
Abhishek et al. (2022)	ResNet50	SVM	2	608	98.00
ALL - AML - CLL - CML - HBS					
Ahmed et al. (2019)	Proposed CNN	CNN	2	903	81.74
Bibi et al. (2020)	CNN DenseNet121	CNN	2	518	99.91

curacy. Vogado et al. (2018) used eight datasets: the ALL-IDB1, ALL-IDB1 (Crop), ALL-IDB2 (Labati et al., 2011), Leukocytes (Sarrafzadeh and Dehnavi, 2015), CellaVision (Rollins-Raval et al., 2012), Atlas (Sarrafzadeh et al., 2014) and (Sarrafzadeh et al., 2015) datasets. The total number of used images was equal to 1,268, and CNNs VGG-F (Chatfield et al., 2014), AlexNet (Krizhevsky et al., 2012) and CaffeNet (Jia et al., 2014) were used to extract the used descriptors; the used classifier was the *support vector machine* (SVM), and an accuracy of 99.76% was achieved. Loey et al. (2020) used the AlexNet network (Krizhevsky et al., 2012) for feature extraction and classification. The used dataset has 564 images: 282 healthy images and 282 images with leukemia, which were obtained from the American Society

of Hematology (ASH) (ASH, 2020) and Kangle[1], and techniques of data augmentation were applied to reduce overfitting. Vogado et al. (2021) developed a network called LeukNet to diagnose 3,536 images, which are also used in this study; these images were gathered from 18 different datasets, and the authors obtained an accuracy of 98.61%.

Singhal and Singh (2016) extracted texture features of images from the ALL-IDB2 dataset (Labati et al., 2011), which includes 130 healthy images and 130 images with ALL, and obtained an accuracy of 93.80%. Shafique and Tehsin (2018) applied the AlexNet convolutional neural network (Krizhevsky et al., 2012) to 368 images from the ALL-IDB1 and ALL-IDB2 datasets (Labati et al., 2011), where 179 images present ALL and 189 are healthy. The authors applied data augmentation and obtained 99.50% of accuracy on the binary classification problem. Ahmed et al. (2019) proposed a new CNN to perform the classification of healthy images and images with ALL. The tests used images from the ASH (ASH, 2020) and ALL-IDB (Labati et al., 2011) datasets; in total, there were 179 images with ALL and 175 healthy images. By also applying techniques of data augmentation based on rotation, translation, flip, shear, and zoom, Ahmed et al. (2019) obtained an accuracy of 88.25%.

As to the second group of approaches, Pansombut et al. (2019) used the ASH (ASH, 2020) and the ALL-IDB1 (Labati et al., 2011) datasets to classify the imaged cells as either ALL or healthy cells. The authors used 121 images, applied techniques of data augmentation, and developed a network called ConvNet, which obtained 81.74% of accuracy. Gehlot et al. (2020) developed a new CNN called SDCT AuxNet$^{\theta}$. The training and testing steps were performed using a private ALL cancer dataset of 118 subjects (Gupta et al., 2019), and an accuracy of 93.40% was obtained for 15,114 segmented cell images.

Das and Meher (2021) developed a hybrid model based on the MobilenetV2 and ResNet18 networks to detect healthy and ALL blood slides. Two datasets (ALL-IDB1 and ALL-IDB2) with 368 images (179 with ALL and 189 healthy) were used. Two tests were performed with a different number of training and test images. The best result was achieved by using 70% of the images for training and 30% for testing. Accuracies of 97.92 and 96.00% were obtained on the ALL-IDB1 and ALL-IDB2 datasets, respectively.

[1]www.kaggle.com/paultimothymooney/blood-cells

Khandekar et al. (2021) used the Yolov4 model to detect the cells and classify them into ALL and healthy cells. Two datasets were used: the ALL-IDB1 and C-NMC-2019 datasets. For the first dataset, the validation was done on 21 images, and 11 were used to test the model. The remaining 76 images were used for training, and after augmenting the 11 test images, the cells belonging to each class were counted and compared to the ground truth labels. As to the C-NMC-2019 dataset, the test was performed on 1000 images, with 500 images belonging to each class. The obtained mean Average Precision (mAP) was equal to 96.06% for the ALL-IDB1 dataset and 98.7% for the C-NMC-2019 dataset, respectively.

Zakir Ullah et al. (2021) used the C-NMC-2019 dataset as the input to the VGG16 network, and applied the efficient channel attention (ECA) module (Wang et al., 2021) in each convolutional block of this network to further increase the relevance of the extracted resources. The authors compared the performances of the VGG16 model with and without the ECA module. The latter showed that this attention mechanism helps to improve the accuracy of the used model, as it explores the relationship between channels and obtains a better representation of the resources. The accuracy obtained was equal to 91.10%.

The research of Karar et al. (2022) proposed an intelligent framework for classifying acute leukemias using blood microscopy images. Former blood samples were collected using digital devices without microscopy and sent to a cloud server. Then, a cloud server used a generative adversarial network (GAN) classifier to automatically identify blood conditions, leucemias, and healthy conditions. The developed classifier was evaluated on two public datasets: ALL-IDB and the ASH image bank. The authors achieved accuracy rates of 98.67% for the binary classification (ALL or healthy) and 95.5% for the multiclass classification (ALL, AML, and healthy blood cells).

Rodrigues et al. (2022) proposed a hybrid model using a genetic algorithm (GA) and a neural convolutional residual network (ResNet-50V2) to predict ALL using microscopy images available in the ALL-IDB dataset. The genetic algorithm was used to find the best hyperparameters leading to the highest precision level in the classification. The results shown that the optimizing with GA improved the classifier precision, which was of 98.46%. Experiments were conducted on cut cells and real-size microscopy images and data augmentation was applied. The proposed model was tested on two datasets: the ALL-IDB1 and ALL-IDB2 datasets. Several tests were performed accroding to two classification scenarios: ALL-HBS and ALL-AML-HBS, with the best

result accuracy obtained for the binary classification scenario equal to 97% when the adjustment technique was applied to the Vgg16 network. The best accuracy obtained for the classification according to three classes was of 95%, which was obtained when the SVM was trained using the ResNet50 features.

Among the tests presented in Rastogi et al. (2022), a binary classifica-tion experiment using the ALL-IDB2 dataset was described, where different trained classifiers were used, with the extra trees classifier (ETC) obtaining the best performance from features extract using a deep network feature extractor developed by the authors: LeuFeatx. The reported accuracy was of 96.15%.

For the third group of approaches, two articles that classify images into AML and HBS images were found. Madhukar et al. (2012) developed a classification system that improved the contrast of the input image and extracted five features from it. The SVM classifier was used on 50 images included in the ASH (ASH, 2020) dataset, and an accuracy of 93.50% was obtained. Goutam and Sailaja (2015) used 90 images from the ASH dataset (ASH, 2020) and attained an accuracy of 98.00%. From the number of images in-cluded in this dataset, one can realize that it is like a reference dataset, and that each work based on it uses a different subset of the included images.

Dasariraju et al. (2020) suggested a method for identification and categorization AML using a random forest classifier. Images of leukocytes in AML patients and healthy controls were obtained from a dataset publicly avail-able in The Cancer Imaging Archive. Sixteen features of each image of white blood cells were extracted, and the five most important features were used in the classification step. The number of images from each kind of leuko-cyte was 1,274. The random forest classifier was trained for the detection and classification of immature leukocytes. The model achieved 92.99% of accuracy for detection, and 93.45% of accuracy for classification of immature leukocytes into four types.

Few studies that detect chronic leukemia were found, and one can believe that this is due to the unavailability of images. Particularly, for the fourth group of approaches, only the work by Khosla and Ramesh (2018), which is based on CNN concepts, was found. Chronic leukemia does not have many publicly available images, so the authors used data augmentation. There were 67 original slides, and the authors divided them into four patches, which were rotated twice to achieve data augmentation, which led to 536 images, and an accuracy of 97.60% was reported.

The fifth group of approaches has the goal of the classification of the

input images into three classes: ALL, AML and HBS. The works found for this group have the main focus on the discovery of acute leukemia. Rawat et al. (2017) performed the segmentation of the leukocyte nucleus in 240 images; there were 100 healthy images, 60 with ALL and 80 with AML. Then, they analyzed 331 characteristics of each segmented core using an SVM classifier. Laosai and Chamnongthai (2018) also took these classes into account, and performed tests on 500 images: 150 of the ALL type, 150 of the AML type and 200 of the HBS type. These images were acquired at the Ubonratchathani Cancer Hospital and Sunpasit International Hospital, in Thailand, and, according to the authors, promising results were achieved. For this type of approaches, the works by Tran et al. (2018) and Claro et al. (2020) were also found. These authors used convolutional neural networks to classify both acute leukemia and non-leukemia images. In the first work, the developed system was called: LeukemiaNet, and in the second one, the proposed network was designated as: Acute Leukemias Recognition Network (Alert Net-RWD). In the first work, 108 images from the ALL-IDB1 dataset (Labati et al., 2011) and 33 AML images gather from the Internet were used, while in the second work, 2,515 images from 16 different datasets were used. In both works, data augmentation was applied, and accuracies of 97.2 and 97.18%, respectively, were obtained.

The last group of approaches involves the classification of images into five classes: ALL, AML, CLL, CML and HBS. The solution proposed by Ahmed et al. (2019) also performed the classification on images with ALL and *HBS*. These authors developed a CNN and applied it to a total of 903 images: 179 ALL, 179 AML, 185 CLL, 185 CML and 175 HBS images, which were gathered from the ASH (ASH, 2020) and ALL-IDB (Labati et al., 2011) datasets. Data augmentation techniques such as rotation, translation, flip, shear and zoom, were applied, which led to an accuracy of 81.74%. Bibi et al. (2020) applied the DenseNet-121 convolutional neural network (Huang et al., 2017) on 518 images: 181 ALL, 55 AML, 38 CLL, 57 CML and 187 HBS images. They also used data augmentation to reduce overfitting and achieved an accuracy of 99.91%.

3. Materials and methods

This study aimed to evaluate the influence of using data augmentation and combinations of CNNs on the detection of leukemia types in blood slide images. The identification of leukemia types in images is a challenging is-

sue. Here, five leukemia classification problems were addressed, mainly three binary classification and two multiclass classification problems: 1) leukemia vs. HBS, 2) ALL vs. HBS, 3) AML vs. HBS, 4) ALL vs. AML vs. HBS, and 5) ALL vs. AML vs. HBS vs. other types. Public image datasets were used for these classes, which were evaluated with and without data augmen-tation, and with CNNs combined into two different ways: a multilevel CNN configuration and a CNN ensemble configuration.

3.1. Proposed evaluation

The evaluation methodology used in this study follows the flowchart shown in Figure 2. Initially, publicly available datasets were searched, as the quality of the used image dataset has a significant impact on the classification performance. Thus, 3,536 images of blood slides were collected. Even though this is a number superior to those used in the literature, it is still insufficient for the adequate training of a CNN (Chen, 2019). Therefore, the solution used in this study was to increase the generalization capacity of the classification models using techniques of data augmentation (DA). The techniques of DA generate new samples for training, and the performances of the developed models were evaluated with and without the use of these techniques. Then, the five scenarios under study for the gathered dataset were took into account, and seven neural networks were used to extract features from the input images and classify them. After performing all these experiments, the multilevel and ensemble configurations were applied and evaluated.

3.2. Image dataset

The development of a robust methodology to aid in diagnosis strongly depends on the data used in its validation. The main challenges found in the reviewed state-of-the-art methods are related to the used datasets, since most of them are private. However, in this study was possible to ob-tain 18 public datasets with 3,536 images for the evaluation of the models under study. The restrictions for using a given dataset in the current re-search were the following: it had to be a public image dataset, it had to have the ground-truth of the images classification, and it had to be used in works found in the literature. Table 2 presents the used image datasets, including information about the number of comprised images per class. The leukemia-images (http://www.leukemia-images.com/ (accessed on 16 August 2021)), UFG dataset (https://hematologia.farmacia.ufg.br (accessed on 16

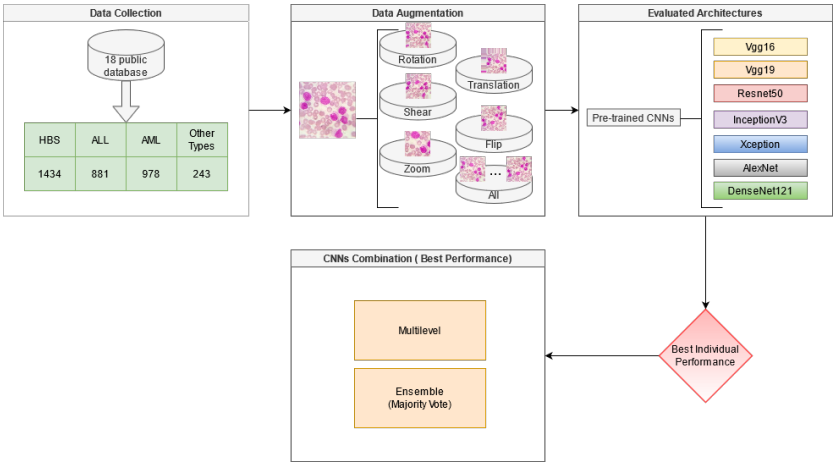


Figure 2: Flowchart of the evaluation methodology used to detect different types of leukemia from images.

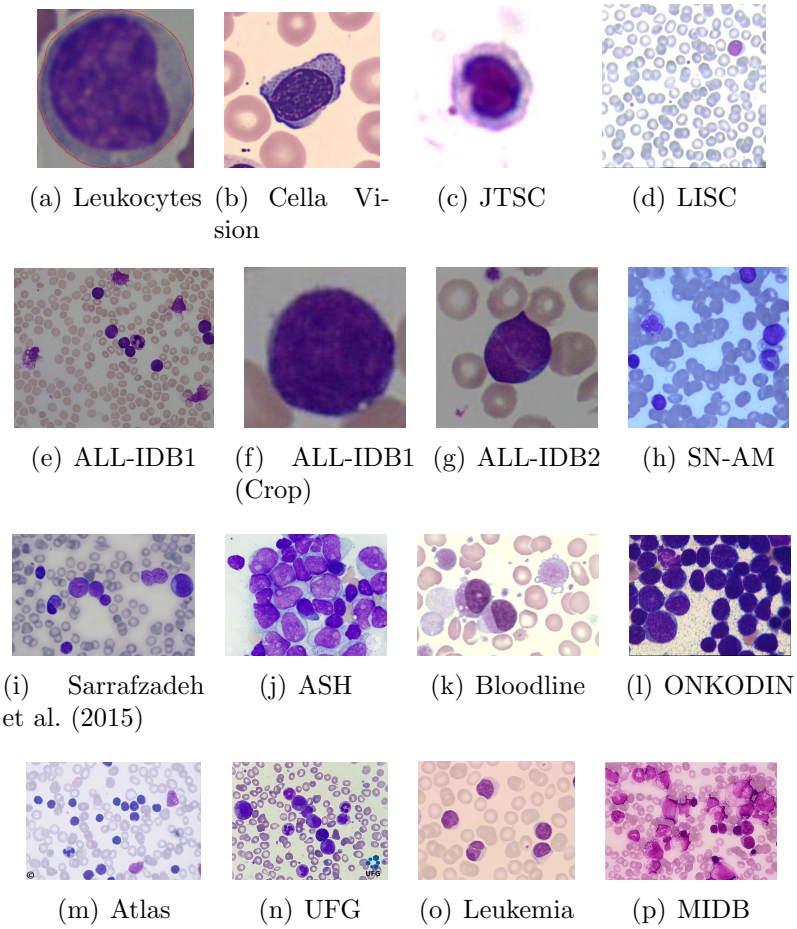
August 2021)) and MIDB (http://www.midb.jp/blood_db/db.php?lang=en (accessed on 16 August 2021)) datasets were obtained from the indicated URLs.

Table 2: Summary of the used image datasets.

Dataset	HBS	ALL	AML	Other types	Total	Size	Focus on cells	Country	Ref.
ALL-IDB 1	59	49	-	-	108	2592 × 1944	No	Italy	Labati et al. (2011)
ALL-IDB 1 (Crop)	-	510	-	-	510	2592 × 1944	Yes	Italy	Labati et al. (2011)
ALL-IDB 2	130	130	-	-	260	257 x 257	Yes	Italy	Labati et al. (2011)
Leukocytes	149	-	-	-	149	609 x 584	Yes	Iran	Sarrafzadeh and Dehnavi (2015)
CellaVision	109	-	-	-	109	399 x 399	Yes	Sweden	Rollins-Raval et al. (2012)
Atlas	-	25	40	23	88	460×307	No	-	-
Omid et al. 2014	154	-	-	-	154	3872×2592	No	Iran	Sarrafzadeh et al. (2014)
Omid et al. 2015	-	-	27	-	27	3246×2448	No	Iran	Sarrafzadeh et al. (2015)
ASH	-	-	96	-	96	184 × 138	No	-	ASH (2020)
Bloodline	-	-	217	-	217	1280 x 720	Yes/No	Brazil	Vale et al. (2014)
ONKODIN	-	-	78	-	78	768 x 576	No	Germany	Böhm (2008)
CellaVision 2	100	-	-	-	100	300×300	Yes	Sweden	Zheng et al. (2018)
JTSC	300	-	-	-	300	2048 × 1536	Yes	China	Zheng et al. (2018)
UFG	57	10	27	27	121	640 x 480	Yes/No	Brazil	link
SN-AM	-	30	-	-	30	224 x 224	No	India	Gupta et al. (2019)
leukemia-images	-	40	78	22	140	755×570	No	-	link
MIDB Dataset	-	87	415	171	673	768×576	No	World	link
LISC Dataset	376	-	-	-	376	720×576	No	Iran	Rezatofighi and Soltanian-Zadeh (2011)
Total of images	**1,434**	**881**	**978**	**243**	**3,536**				

From the number of images indicated in Table 2, one can verify that, out of the total number of images, 1,434 images belong to the HBS class (40.55% of the total), 881 belong to the ALL class (24.92%), 978 belong to the AML class (27.66%) and 243 belong to the "other types" class (6.87%). Hence, it can be realized that the gathered image dataset is unbalanced.

Some of the 18 gathered datasets have images with only one leukocyte per image, and others have multiple leukocytes per image. Only the UFG and Bloodline datasets have both kinds of images. Hence, the combination of the gathered datasets resulted in a dataset of images with distinct color, texture, contrast and resolution. This diversity created a challenge for the models under study. Figure 3 shows examples of the used images.



(a) Leukocytes (b) Cella Vision (c) JTSC (d) LISC

(e) ALL-IDB1 (f) ALL-IDB1 (Crop) (g) ALL-IDB2 (h) SN-AM

(i) Sarrafzadeh et al. (2015) (j) ASH (k) Bloodline (l) ONKODIN

(m) Atlas (n) UFG (o) Leukemia (p) MIDB

Figure 3: Examples of the used (a-d) HBS images, (e-h) ALL images, (i-l) AML images and (m-p) Other type of images.

3.3. Data augmentation

Overfitting occurs when a model learns the details, including the existent noise, in the training data, which means that noise or random fluctuations

in the training data are picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data, and therefore, negatively impact the ability of the model to generalize (Shorten and Khoshgoftaar, 2019).

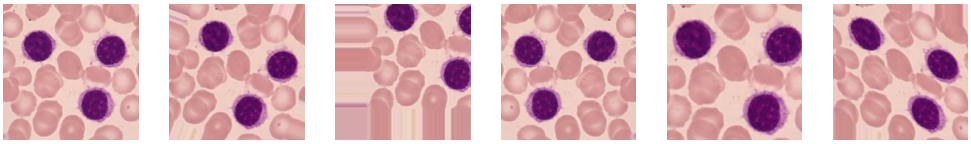
There are several strategies that have been proposed for overfitting reduction. Most of the strategies that contribute to increase the generalization capability are focused on the architecture of the model used. Hence, there are strategies that aim to improve deep learning models for applications with smaller datasets, such as dropout, batch normalization, transfer learning and pre-training (Kukačka et al., 2017). Contrary to these techniques, data augmentation attempts to solve overfitting by solving the problem of a limited available training dataset.

Limited data is a significant obstacle to the application of deep learning models. Moreover, unbalanced classes can be an additional hurdle to tackle. Although there may be enough data for some equally essential classes, classes with reduced data will suffer from a low class-specific accuracy (Shorten and Khoshgoftaar, 2019).

Image augmentation is essential in deep learning-based methods, regardless of the problem under study Li et al. (2021). Particularly, in medicine, the use of image augmentation is necessary to successfully achieve the segmentation, classification and analysis of images with deep learning-based models. There are articles in the literature about techniques of data augmentation for specific types of images, such as skin images, as, for example, (Goceri, 2020). The literature also presents several different techniques of data augmentation, including based on image semantics (Wang et al., 2021). However, these methods need the definition of several configuration parameters. Thus, it was decided in this study to apply the most common techniques of data augmentation, which are based on geometric transformations such as the rotation, translation, flipping, scale and shear transformations (Figure 4).

3.4. Evaluated architectures

According to Sarvamangala and Kulkarni (2021), several CNNs have been proposed to classify medical images. However, most authors used CNNs that were pre-trained on the ImageNet challenge dataset. However, unfortunately, it is not always straightforward to reproduce the study of related articles. Hence, to perform the experiments under study, pre-trained CNNs that have attained good performance on the ImageNet challenge dataset, particularly,



(a) (b) (c) (d) (e) (f)

Figure 4: Examples of data augmentation transformations: (a) the original image, and the image after (b) a rotation of 40º, (c) a translation of 20%, (d) flipping, (e) a scaling by a factor of 20% and (f) a shear of 20º.

the AlexNet, VGG (VGG16 and VGG19), ResNet (ResNet50), GoogLeNet (InceptionV3 and Xception), and DenseNet121 networks, were evaluated.

The network Alexnet, which was developed by Krizhevsky et al. (2012), was designed for the ILSVRC-2010 competition, mainly to carry out training and classification on the ImageNet dataset. It comprises eight layers that need to be trained: there are five convolutional layers with filters of 5×5 and 7×7 kernels, which are followed by three fully connected layers and max-pooling layers.

Simonyan and Zisserman (2014) introduced the VGG network architecture, which is characterized by its simplicity: it uses only 3×3 convolutional layers stacked on top of each other with increasing depth. The reducing volume size is handled by max pooling. Two fully connected layers, each with 4,096 nodes, are followed by a softmax classifier, and "16" and "19" stand for the number of weight layers in the network (Simonyan and Zisserman, 2014).

ResNet50 (He et al., 2016) is a deep convolutional network architecture proposed in 2016 to solve the vanishing gradient problem, which causes saturation in learning and, consequently, slows down the training. The basic idea is to skip blocks of convolutional layers using shortcut connections to form unions called residual blocks, which significantly improve the training efficiency, and mostly solve the degradation problem generally present in deep networks.

The InceptionV3 (Szegedy et al., 2016) architecture emerged as a new version of the GoogLeNet and InceptionV2 architectures. This architecture reduces the complexity of a CNN in terms of the number of operations performed using Inception modules, which consist of parallel combinations of layers with convolutional filters of size 1×1, 3×3, and 5×5. Convolutions with larger filters are computationally more costly; therefore, the authors

proposed performing 1×1 convolutions first, reducing the dimensionality of the characteristics map, and then performing convolutions with the other filters. The Inception modules result in a reduction of 28% in the number of parameters relatively to traditional convolutional layers.

The Xception (Chollet, 2017) architecture is an extension of the Inception architecture that replaces the standard Inception modules with separable convolutions in depth. Instead of partitioning the input data into multiple compressed blocks, it maps the spatial correlations for each output channel separately. Then, it performs a convolution of 1×1 in-depth to capture the correlation between channels. This operation is essentially equivalent to an existing process known as depthwise separable convolution, which consists of a depthwise convolution - a spatial convolution performed independently for each channel, followed by a pointwise convolution, using filters with size 1×1 between channels. Xception achieved superior results compared to previous versions, despite having fewer layers and parameters. The inclusion of the depthwise separable convolution layers also provided greater efficiency in terms of the computational cost; it is less costly and faster than the standard convolution by performing less operations.

The dense convolutional network, or DenseNet, is a dense block used to improve the flow of information between layers. This network uses fewer parameters than ResNet for its training (Huang et al., 2017). DenseNet121 consists of 121 layers, and each layer is connected to all subsequent layers. In addition, each layer receives important features learned by any previous layers of the network, which makes the network training more efficient (Li et al., 2018).

3.5. Multilevel configuration

Due to the variety of sizes and morphologies of white blood cells, the extraction of various features using different CNNs is interesting. To ensure that helpful information is not lost, in this study, a CNN based on the multilevel concept was evaluated. The main idea behind using this concept is that each used CNN will extract different features and, at the end, all these features are concatenated.

Hence, according to Lyu and Ling (2018), all feature maps are concatenated into a one-dimensional vector and then connected to a fully connected layer, as show in Figure 5. This strategy can help to improve the accuracy of the classification model.

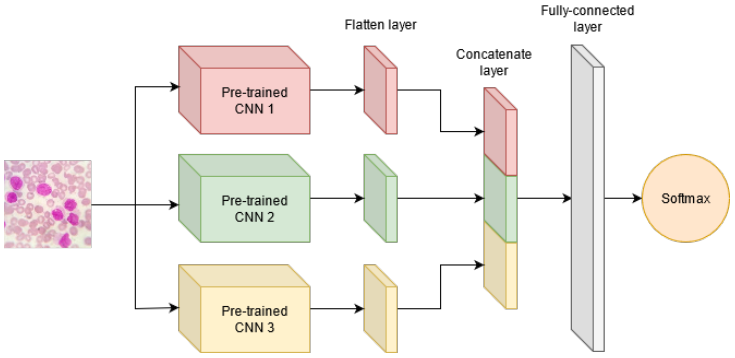


Figure 5: Generic scheme showing how a CNN multilevel configuration works.

This adopted multilevel setup is based on a few main ideas. First, all feature maps from all CNNs are flattened and concatenated into a one-dimensional vector, and then connected with a fully connected layer. This strategy can help to improve the classification accuracy of the classification model, as each CNN plays an important role in the final result, since they extract various features to represent the input image, whether are focused on the cells or on the blood slide images. Second, this technique helps to improve the variance rate of the classification results. Variance refers to the sensitivity of the learning model to the specifics of the training data, e.g., the noise and specific observations. This is interesting, as the model will be specialized to the data by learning random noise, and will be varied each time it is trained on different data (Belkin et al., 2019).

3.6. Ensemble technique

The ensemble technique was proposed long before the emergence of the deep learning paradigm (Dasarathy and Sheela, 1979). The theory behind this technique is quite simple and supported by the well-known notion of "the wisdom of crowds": instead of relying on just one model for prediction, a set of multiple pre-trained models is created; then, the results of the models are combined into a final classification by merging their votes. The original idea was developed to reduce the variance of the classifiers in order to achieve a better overall performance (Dietterich, 2000).

The majority vote method is one of the most popular methods used in ensemble based classifiers. This method involves taking the decision of each model and selecting the one with the most votes as the final decision. That is, the class label provided for a specific sample, i.e., prediction, will be the class

label representing the majority of the class labels supplied for each classifier (Dietterich, 2000).

As our problem has four classes, a tie may occur when the majority voting technique is used. In this case, the accuracy values achieved by the CNNs during the training stage are used as a tiebreaker criterion. For example, if each CNN in the ensemble predicts a different class, the final decision will be that of the CNN that obtained the highest accuracy in the training step. Figure 6 shows how the adopted ensemble technique works. Particularly, this figure presents an example output for four elements from a dataset. There was no tie in the first three cases: element 1 belongs to class 1, element 2 to class 0 and element 3 to class 3. However, in the case of the fourth element, each CNN voted for a different class, creating a tie. The final result was class 0 (zero), as this class was the output of CNN 3, the network that obtained the best accuracy during the training.

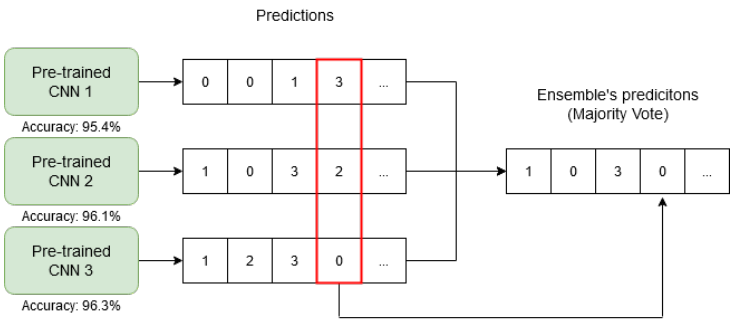


Figure 6: Example of the adopted ensemble classification method.

3.7. Evaluation metrics

To analyze the classification results, the confusion matrix was computed. Then, from the elements of this matrix, the accuracy (A), precision (P), recall (R) and F1-*score* were calculated (Powers, 2007).

The kappa index (k), which is recommended as an appropriate exactitude measure because it can adequately represent the confusion matrix, was also computed. This index takes all elements of the confusion matrix into account, rather than just those on the main diagonal; the global classification accuracy only considers the main diagonal elements. It can be calculated as:

$$\mathrm{k} = \frac{observed - expected}{1 - expected}. \tag{1}$$

According to Landis and Koch (1977), k assumes values between 0 (zero) and 1 (one). The result is classified according to the k value as follows: k $\leq$ 0.2: Bad; 0.2 $<$ k $\leq$ 0.4: Fair; 0.4 $<$ k $\leq$ 0.6: Good; 0.6 $<$ k $\leq$ 0.8: Very Good and k $>$ 0.8: Excellent.

The cost function metric (loss) was also used in this study. This function is responsible for showing how far one is from the ideal prediction, and, therefore, it quantifies the "cost" or "loss" by accepting the prediction generated by the current parameters of the model (Janocha and Czarnecki, 2017).

4. Experiments and results

The dataset used to study the five scenarios under evaluation is composed of the following images: 1,434 images of healthy slides, 881 images of ALL, 978 images of AML and 243 images of "other types" of leukemia. K-fold cross-validation with the value of k equal to 5 was applied in the evaluated experiments, which were performed on a PC with a 3.6 GHz Intel®Xeon™processor with 24GB of RAM and an NVIDIA TITAN XP 12GB graphics card.

The influence of the use of data augmentation on the classification results was first investigated. To do this, the results obtained by the seven networks introduced in Section 3.4 with and without the use of DA were compared. For the last scenario, it was also studied the individual results for each used data augmentation technique.

4.1. Binary classification

The binary classification problem was subdivided into three sub-problems: leukemia vs. HBS, ALL vs. HBS, and AML vs. HBS. These sub-problems were chosen and analyzed due to two main reasons: the first one was the availability of studies in the literature that carry out the same classification, and the second reason was the availability of public datasets for each class.

4.1.1. Leukemia vs. HBS

For this scenario, images from the 18 gathered datasets were used (3,536 images in total), 1,434 from the HBS class and 2,102 from the leukemia class. Table 3 presents the results obtained using the seven neural networks under study with and without DA. Under this scenario, ResNet50-DA obtained the best outcome in terms of all of the used classification metrics. One can see that the AlexNet network obtained a good performance, but its results were inferior compared to the ones obtained by the other networks.

Table 3: Results obtained for the leukemia vs. HBS scenario, with and without data augmentation (DA), after applying k-fold cross-validation. (The best values are in bold.)

Model	A(%)	P(%)	R(%)	F1-score(%)	K
AlexNet	95.87±0.51	96.09±1.12	96.99±1.34	96.53±0.44	0.9141±0.010
AlexNet-DA	95.25±0.81	95.38±1.41	96.70±2.15	96.01±0.71	0.9011±0.016
VGG16	99.15±0.22	99.00±0.30	99.56±0.35	99.28±0.18	0.9823±0.004
VGG16-DA	99.23±0.29	99.04±0.28	99.66±0.27	99.35±0.24	0.9841±0.006
VGG19	98.98±0.36	99.09±0.31	99.18±0.40	99.13±0.31	0.9788±0.007
VGG19-DA	99.23±0.15	99.18±0.12	99.51±0.28	99.35±0.13	0.9841±0.003
ResNet50	99.43±0.26	99.37±0.36	99.66±0.13	99.52±0.22	0.9882±0.005
ResNet50-DA	**99.54**±0.27	**99.47**±0.53	**99.75**±0.23	**99.61**±0.23	**0.9904**±0.005
InceptionV3	98.58±0.36	98.89±0.54	98.71±0.31	98.80±0.30	0.9706±0.007
InceptionV3-DA	99.23±0.15	99.28±0.35	99.40±0.13	99.34±0.13	0.9838±0.003
Xception	98.19±0.68	98.19±0.63	98.75±0.51	98.47±0.57	0.9623±0.014
Xception-DA	98.16±0.51	98.56±0.66	98.32±0.47	98.44±0.42	0.9618±0.010
DenseNet121	99.32±0.27	99.28±0.16	99.56±0.39	99.42±0.23	0.9858±0.005
DenseNet121-DA	99.37±0.16	99.28±0.28	99.66±0.27	99.47±0.13	0.9870±0.003

4.1.2. ALL vs. HBS

There is a good number of articles in the literature addressing this scenario, as there are many public datasets available for it. Here, 14 datasets with 2,315 images in total: 1,434 healthy images and 881 images with ALL, were used. Table 4 indicates the performance obtained by the seven networks under study with and without DA. The data presented in this table shows that the DenseNet121-DA network obtained the best result in terms of accuracy, recall, F1-score and kappa index. On the other hand, the ResNet50 network obtained the best accuracy (99.71%).

4.1.3. AML vs. HBS

For the last binary classification scenario, 2,412 images from 16 datasets were used, including 1,434 healthy images and 978 images with AML. Table 5 presents the performances obtained by each of the seven networks under study with and without DA. The data presented in this table show that the ResNet50-DA network obtained the best result in terms of accuracy, precision, F1-score, and kappa index. On the other hand, the DenseNet121 network had the best recall value (99.71%).

From the obtained results, one can see that an improvement was achieved by applying data augmentation in practically all evaluated networks for the three binary classification scenarios under study.

Table 4: Results obtained for the ALL vs. HBS scenario, with and without data augmentation (DA), after applying k-fold cross-validation. (The best values are in bold.)

Model	A(%)	P(%)	R(%)	F1-score(%)	K
AlexNet	96.19±0.69	96.74±0.94	97.14±0.96	96.93±0.55	0.9192±0.014
AlexNet-DA	97.32±0.69	97.57±1.10	97.81±1.42	97.68±0.56	0.9392±0.014
VGG16	99.35±0.43	99.44±0.52	99.51±0.31	99.47±0.34	0.9862±0.009
VGG16-DA	99.52±0.18	99.71±0.28	99.50±0.30	99.61±0.14	0.9898±0.003
VGG19	99.08±0.21	99.30±0.35	99.30±0.35	99.29±0.17	0.9816±0.004
VGG19-DA	99.57±0.30	99.78±0.19	99.51±0.39	99.65±0.24	0.9908±0.006
ResNet50	99.61±0.28	99.65±0.24	99.71±0.29	99.68±0.22	0.9917±0.005
ResNet50-DA	99.74±0.38	**99.79**±0.31	99.79±0.31	99.79±0.31	**0.9944**±0.008
InceptionV3	99.30±0.41	99.23±0.56	99.64±0.24	99.40±0.33	0.9852±0.008
InceptionV3-DA	99.43±0.12	99.51±0.39	99.57±0.38	99.54±0.09	0.9880±0.002
Xception	98.79±0.58	98.54±0.56	99.50±0.39	99.02±0.47	0.9742±0.012
Xception-DA	98.70±0.48	98.61±0.53	99.30±0.35	98.95±0.38	0.9696±0.011
DenseNet121	99.69±0.24	99.65±0.42	**99.85**±0.18	99.75±0.19	0.9935±0.005
DenseNet121-DA	**99.78**±0.26	99.78±0.31	**99.85**±0.18	**99.82**±0.21	**0.9954**±0.005

4.2. Multiclass classification

For multiclass classification, two scenarios were studied: The first scenario involved dividing the images into the ALL vs. AML vs. HBS classes, and the second scenario involved the ALL vs. AML vs. HBS vs. other type images. This last scenario includes the "other types" class due to the small number of images available for some types of leukemia, such as chronic leukemia. Thus, if this class was divided further, it would be even more unbalanced. So far, any work that performs this division has been found; therefore, more tests for this scenario were performed in this study.

4.2.1. ALL vs. AML vs. HBS

For this scenario, a total of 3,293 images from the 18 gathered datasets, including 1,434 images from the HBS class, 881 from the ALL class and 978 from the AML class, were used. One can find the results obtained for this scenario in Table 6. Relatively to the results obtained for the binary classification problem, the results presented in this table indicate a reduced performance. The DenseNet121-DA network achieved the best performance, with 97.11% of accuracy.

4.2.2. ALL vs. AML vs. HBS vs. other types

The dataset used in developing this scenario is composed of 1,434 images of healthy slides, 881 images of ALL, 978 images of AML and 243 images of "other types" of leukemia. As with the other scenarios, the use of data

Table 5: Results obtained for the AML vs. HBS scenario, with and without data augmentation (DA), after applying k-fold cross-validation. (The best values are in bold.)

Model	A(%)	P(%)	R(%)	F1-score(%)	K
AlexNet	96.81±0.47	97.48±0.78	97.13±1.08	97.30±0.41	0.9338±0.009
AlexNet-DA	96.93±0.71	96.78±0.97	98.11±1.53	97.43±0.60	0.9362±0.014
VGG16	99.17±0.29	99.23±0.28	99.50±0.39	99.30±0.24	0.9827±0.006
VGG16-DA	99.66±0.18	99.71±0.15	**99.71**±0.28	99.71±0.15	0.9931±0.003
VGG19	99.04±0.45	99.09±0.39	99.29±0.65	99.19±0.38	0.9802±0.009
VGG19-DA	99.42±0.22	99.64±0.24	99.36±0.51	99.50±0.19	0.9879±0.004
ResNet50	99.46±0.47	99.64±0.34	99.43±0.52	99.54±0.40	0.9888±0.009
ResNet50-DA	**99.70**±0.18	**99.92**±0.15	99.57±0.38	**99.75**±0.15	**0.9939**±0.003
InceptionV3	98.59±0.17	98.27±0.47	99.30±0.24	98.78±0.17	0.9698±0.004
InceptionV3-DA	99.46±0.23	99.43±0.30	99.64±0.24	99.54±0.19	0.9887±0.004
Xception	98.05±0.63	97.73±0.81	99.02±0.29	98.37±0.52	0.9594±0.013
Xception-DA	97.80±0.27	97.13±0.48	99.23±0.57	98.16±0.23	0.9541±0.005
DenseNet121	99.42±0.30	99.64±0.24	99.37±0.45	99.50±0.26	0.9879±0.006
DenseNet121-DA	99.67±0.10	99.65±0.15	**99.71**±0.28	99.71±0.09	0.9931±0.002

augmentation was evaluated. Table 7 presents the results of this evaluation, where one can see an improvement in terms of the used performance metrics in all the networks evaluated with the use of DA. The DenseNet121 architecture obtained the best performance, with accuracy, precision, recall and F1-score values equal to or greater than 94%.

Each data augmentation technique can have a different impact in terms of the overall robustness it brings to a neural network. It is known that some techniques of data augmentation improve the decision stability of the used neural network, but others can have a detrimental effect (Shorten and Khoshgoftaar, 2019). Based on this and the obtained results, individual tests were performed for each technique of DA with the seven network architectures under study. Each network was trained according to the five scenarios under study, and five data augmentation transformations were used: rotation, translation, flip, zoom and shear.

Table 8 indicates the best performance obtained among the five scenarios for each network. From this table, one can confirm that the best results for ResNet50 and DenseNet121 were found when rotation was applied as a data augmentation technique. On the other hand, InceptionV3 and AlexNet had the best performance when translation was applied. For the Xception, VGG16 and VGG19 networks, the flip operation led to the best performance. Analyzing the results in Tables 7 and 8, one can verify that the DenseNet121 network with data augmentation obtained the best performance, achieving an accuracy of 94%.

Table 6: Results obtained for the ALL vs. AML vs. HBS scenario, with and without data augmentation (DA), after applying k-fold cross-validation. (The best values are in bold.)

Model	A(%)	P(%)	R(%)	F1-score(%)	K
AlexNet	89.58±1.44	89.93±1.23	89.58±1.44	89.50±1.51	0.8394±0.022
AlexNet-DA	89.58±1.69	90.00±1.38	89.71±1.74	89.42±1.88	0.8389±0.026
VGG16	96.41±0.69	96.49±0.71	96.42±0.69	96.41±0.70	0.9448±0.010
VGG16-DA	96.99±0.82	97.01±0.84	97.00±0.82	96.99±0.81	0.9537±0.012
VGG19	96.05±0.55	96.09±0.55	96.05±0.54	96.04±0.56	0.9392±0.008
VGG19-DA	96.81±0.56	96.86±0.59	96.82±0.58	96.80±0.57	0.9509±0.008
ResNet50	95.65±1.07	95.81±1.03	95.72±1.05	95.64±1.09	0.9331±0.016
ResNet50-DA	96.69±0.46	96.70±0.44	96.69±0.45	96.68±0.45	0.9490±0.007
InceptionV3	94.50±0.76	94.66±0.64	94.52±0.76	94.47±0.79	0.9152±0.011
InceptionV3-DA	96.51±0.56	96.58±0.49	96.52±0.53	96.49±0.58	0.9462±0.008
Xception	93.56±0.97	93.70±0.91	93.57±0.95	93.50±1.00	0.9006±0.015
Xception-DA	93.50±1.52	93.65±1.40	93.55±1.45	93.45±1.56	0.8997±0.023
DenseNet121	96.05±0.91	96.15±0.89	96.09±0.89	96.04±0.93	0.9434±0.006
DenseNet121-DA	**97.11**±0.53	**97.13**±0.55	**97.12**±0.55	**97.11**±0.54	**0.9556**±0.008

4.3. Multilevel and ensemble configurations

The second step was to evaluate whether combining the CNNs would benefit the classification performance. Thus, multilevel CNNs and ensembles of CNNs were implemented. As explained earlier, for multilevel CNNs, the flattened layers of the pre-trained networks were concatenated in order to form a new network with various features. The multilevel architecture was evaluated with a configuration of two networks. Then, based on the results obtained in each scenario for each network, they were combined. Therefore, the leukemia vs. HBS scenario included tests with a combination of two networks, and the same was done for the other scenarios. Table 9 indicates the studied combinations of networks that led to the best results for each scenario under study.

Comparing the results presented in Table 9 with the individual results obtained by each CNN presented in the previous tables, one can see that the multilevel technique improved the results only in the multiclass classification problems. Investigating the reason for this, it was found the work of Srinivas et al. (2015), where it is claimed that without the use of combination techniques, together with the use of unbalanced datasets, poor classification performances are generally obtained. This was the case in the multiclass problems studied here, because there are unbalanced classes, mainly for the last scenario. Thus, according to the authors, the use of a multilevel classification approach can address the problems of data imbalance and reduce the variation of the estimation errors.

Table 7: Results obtained for the ALL vs. AML vs. HBS vs. other types classification problem, with and without data augmentation (DA), after applying k-fold cross-validation. (The best values are in bold.)

Model	**A(%)**	**P(%)**	**R(%)**	**F1-score(%)**	**K**
AlexNet	82.74±0.89	82.14±1.16	82.76±0.89	81.46±0.91	0.7461±0.013
AlexNet-DA	84.46±1.05	84.29±1.55	84.28±1.16	83.07±1.30	0.7713±0.016
VGG16	92.72±1.44	92.87±1.42	92.73±1.49	92.72±1.41	0.8946±0.020
VGG16-DA	92.95±0.93	92.98±0.74	92.94±0.93	92.88±0.87	0.8979±0.013
VGG19	91.30±1.02	91.34±1.09	91.32±1.09	91.21±1.09	0.8741±0.014
VGG19-DA	91.65±1.41	91.66±1.50	91.62±1.45	91.59±1.43	0.8792±0.020
ResNet50	91.62±1.25	91.68±1.03	91.63±1.40	91.59±1.20	0.8786±0.018
ResNet50-DA	92.89±0.86	92.87±0.90	92.88±1.11	92.86±0.89	0.8972±0.012
InceptionV3	89.24±1.19	89.12±1.12	89.24±1.09	88.94±1.21	0.8419±0.017
InceptionV3-DA	92.75±0.78	92.69±0.73	92.73±0.90	92.66±0.75	0.8949±0.011
Xception	89.04±1.09	88.86±1.05	89.06±1.48	88.63±1.10	0.8401±0.016
Xception-DA	89.30±1.12	89.18±1.34	89.31±1.12	88.63±1.03	0.8371±0.012
DenseNet121	92.55±1.03	92.78±0.77	92.61±0.97	92.56±0.98	0.8922± 0.014
DenseNet121-DA	**94.00**±0.81	**94.09**±0.82	**94.07**±0.80	**94.08**±0.82	**0.9133**±0.011

Table 8: Results obtained with k-fold cross-validation for the best results achieved by each data augmentation transformation. (The best values are in bold.)

Model	A(%)	P(%)	R(%)	F1-score(%)	K
ResNet50-Rotation	92.78±0.68	92.82±0.66	92.78±0.69	92.74±0.60	0.8954±0.009
InceptionV3-Translation	91.14±0.58	91.27±0.44	91.14±0.52	91.14±0.49	0.8718±0.014
Xception-Flip	89.41±1.01	89.35±0.91	89.41±0.96	88.98±0.96	0.8454±0.014
VGG16-Flip	93.17±1.71	93.16±1.89	93.17±1.70	93.07±1.81	0.9010±0.025
VGG19-Flip	92.21±0.78	92.19±0.66	92.21±0.80	92.08±0.75	0.8871±0.011
AlexNet-Translation	84.69±0.73	83.65±2.09	84.13±1.02	83.13±0.88	0.7746±0.010
DenseNet121-Rotation	**93.69**±0.66	**93.75**±0.58	**93.70**±0.62	**93.70**±0.63	**0.9087**±0.009

In the conducted experiments, the ensemble method was also applied and evaluated. As with the multilevel technique, the previous results were analyzed to built an ensemble in order to improve the classification performance. Thus, the networks that obtained the best performances were selected and combinations of pre-trained CNNs were done as inputs for ensemble models composed of three, four, five and six networks. Table 10 presents the results found in the tests performed for each scenario under study. For all five scenarios, the performance was improved, both in terms of the mean and standard deviation values.

It is accepted that selecting diversified CNNs that present high precision rates in several regions in the characteristics space is essential to building effective multilevel and ensemble configurations. For this reason, model combi-

Table 9: Results of k-fold cross-validation performed on the best classification results obtained using multilevel CNNs. (The best values are in bold.)

Model-multilevel	A(%)	P(%)	R(%)	F1-score(%)	K
Leukemia - HBS					
DenseNet121-DA and VGG19-DA	**99.39**±0.09	**99.28**±0.23	**99.75**±0.16	**99.52**±0.08	**0.9882**±0.002
DenseNet121-DA and VGG16-DA	99.35±0.12	99.27±0.37	99.61±0.31	99.44±0.10	0.9864±0.003
DenseNet121-DA and InceptionV3-DA	99.31±0.16	99.24±0.21	99.59±0.14	99.40±0.10	0.9852±0.003
ALL - HBS					
DenseNet121-DA and VGG16-DA	**99.70**±0.24	**99.85**±0.30	**99.65**±0.24	**99.75**±0.19	**0.9935**±0.005
DenseNet121-DA and InceptionV3-DA	99.61±0.25	99.72±0.28	99.62±0.21	99.68±0.29	0.9913±0.004
ResNet50-DA and Vgg16-DA	99.69±0.25	99.73±0.31	99.63±0.23	99.74±0.20	0.9932±0.006
AML - HBS					
ResNet50-DA and VGG16-DA	**99.70**±0.16	**99.82**±0.15	**99.64**±0.21	**99.75**±0.14	**0.9939**±0.003
DenseNet121-DA and VGG16-DA	99.67±0.18	99.92±0.15	99.51±0.19	99.71±0.15	0.9931±0.004
DenseNet121-DA and Vgg19-DA	99.59±0.14	99.72±0.19	99.58±0.23	99.63±0.24	0.9936±0.004
ALL - AML - HBS					
DenseNet121-DA and VGG19-DA	**97.60**±0.39	**97.74**±0.56	**97.59**±0.38	**97.59**±0.39	**0.9631**±0.006
DenseNet121-DA and VGG16-DA	97.48±0.57	97.49±0.59	97.48±0.57	97.47±0.59	0.9612±0.008
DenseNet121-DA and InceptionV3-DA	97.20±0.68	97.53±0.54	97.21±0.62	97.47±0.39	0.9601±0.007
ALL - AML - HBS - Other types					
DenseNet121-DA and InceptionV3-DA	94.57±0.32	94.59±0.36	94.56±0.32	94.55±0.33	0.9214±0.004
DenseNet121-DA and VGG16-Flip	**94.73**±0.61	**94.75**±0.60	**94.72**±0.62	**94.71**±0.62	**0.9237**±0.008
DenseNet121-DA and VGG19-DA	94.31±0.79	94.37±0.77	94.30±0.79	94.32±0.32	0.9177±0.011

nations were evaluated and it was found the best combination to use in order to build the proposed ensemble configuration. Combining CNN models not only increases the performance, but also reduces the risk of overfitting (Sollich and Krogh, 1995). Thus, a more generalized evaluation was performed. Experimental results are statistically significant for a given level of statistical significance, if they are not attributed to chance and if there is a relationship between results. Thus, one can realize that the proposed approach can help in developing clinically valuable solutions to detect and differentiate leukemia in blood slide images.

Figure 7 depicts the best results found for the experiments with individual CNNs, multilevel CNNs and ensemble configurations for the five scenarios under study. Based on this figure, one can realize the improvements achieved using the ensemble technique relatively to the best individual results of the networks or the multilevel model. One can also realize that for the multiclass problems (Figures 7(d) and 7(e)), the combination models obtained a significant improvement over the best individual performance of the CNN models. Another point to be noted is the individual performance of the DenseNet121-DA and ResNet50-DA models, which obtained the best results in all five scenarios relative to the other evaluated models.

Table 10: Results obtained by the ensemble models. (Best values in bold.)

Model-ensemble	A(%)	P(%)	R(%)	F1-score(%)	K
Leukemia - HBS					
DenseNet121-DA, ResNet50-DA and ResNet50	**99.54**±0.23	**99.54**±0.23	**99.54**±0.23	**99.54**±0.23	**0.9905**±0.004
DenseNet121-DA, ResNet50-DA, DenseNet121 and ResNet50	**99.54**±0.23	**99.54**±0.23	**99.54**±0.23	**99.54**±0.23	**0.9905**±0.004
DenseNet121-DA, ResNet50-DA, DenseNet121, ResNet50 and VGG16-DA	99.43±0.14	99.43±0.14	99.43±0.14	99.43±0.14	0.9882±0.002
DenseNet121-DA, ResNet50-DA, DenseNet121, ResNet50, VGG16-DA and VGG19-DA	99.45±0.11	99.45±0.12	99.46±0.11	99.45±0.11	0.9887±0.002
ALL - HBS					
DenseNet121-DA, ResNet50-DA and DenseNet121	99.82±0.28	99.82±0.27	99.82±0.28	99.82±0.28	0.9963±0.006
DenseNet121-DA, ResNet50-DA, DenseNet121 and ResNet50	99.82±0.28	99.82±0.28	99.82±0.28	99.82±0.28	0.9963±0.006
DenseNet121-DA, ResNet50-DA, DenseNet121, ResNet50 and VGG19-DA	99.91±0.12	99.91±0.12	99.90±0.13	99.91±0.12	0.9981±0.002
DenseNet121-DA, ResNet50-DA, DenseNet121, ResNet50, VGG19-DA and VGG16-DA	**99.95**±0.09	**99.95**±0.09	**99.95**±0.10	**99.95**±0.09	**0.9990**±0.002
AML - HBS					
DenseNet121-DA, ResNet50-DA and VGG16-DA	99.79±0.14	99.79±0.14	99.79±0.14	99.79±0.14	0.9957±0.003
DenseNet121-DA, ResNet50-DA, InceptionV3-DA and VGG16-DA	**99.87**±0.11	**99.87**±0.11	**99.87**±0.11	**99.87**±0.11	**0.9974**±0.002
DenseNet121-DA, ResNet50-DA, ResNet50, InceptionV3-DA and VGG16-DA	99.74±0.09	99.74±0.09	99.74±0.09	99.74±0.09	0.9948±0.002
DenseNet121-DA, ResNet50-DA, ResNet50, InceptionV3-DA, VGG16-DA and VGG19-DA	**99.87**±0.11	**99.87**±0.11	**99.87**±0.11	**99.87**±0.11	**0.9974**±0.002
ALL - AML - HBS					
DenseNet121-DA, VGG16-DA and VGG19-DA	97.35±0.57	97.44±0.53	97.36±0.57	97.35±0.56	0.9593±0.008
DenseNet121-DA, ResNet50-DA, VGG16-DA and VGG19-DA	97.35±0.69	97.41±0.67	97.35±0.69	97.35±0.68	0.9593±0.010
DenseNet121-DA, ResNet50-DA, VGG16-DA, VGG19-DA and InceptionV3-DA	97.44±0.84	97.54±0.76	97.45±0.83	97.44±0.83	0.9607±0.012
DenseNet121-DA, ResNet50-DA, VGG16-DA, VGG19-DA, InceptionV3-DA and VGG16	**97.53**±0.62	**97.61**±0.62	**97.54**±0.62	**97.53**±0.62	**0.9621**±0.009
ALL - AML - HBS - Other types					
DenseNet121-DA, DenseNet121-Translation and DenseNet121-Rotation	94.42±0.30	94.40±0.34	94.41±0.32	94.42±0.32	0.9194±0.004
DenseNet121-DA, DenseNet121-Translation, DenseNet121-Flip and DenseNet121-Rotation	93.82±0.47	94.02±0.47	93.89±0.46	93.86±0.49	0.9104±0.006
DenseNet121-DA, DenseNet121-Rotation, DenseNet121-Translation, DenseNet121-Flip and DenseNet121-Zoom	**94.59**±0.77	**94.65**±0.72	**94.60**±0.78	**94.58**±0.79	**0.9218**±0.011
DenseNet121-DA, DenseNet121-Rotation, DenseNet121-Translation, DenseNet121-Flip, DenseNet121-Zoom and VGG16-Flip	94.45±0.83	94.54±0.77	94.48±0.81	94.45±0.83	0.9196±0.012

5. Discussion

Table 11 allows a comparison among related state-of-the-art methods regarding the addressed classification problem, used number of datasets, used number of images and achieved accuracy. The obtained results suggest that even using general-purpose CNNs, by choosing suitable techniques of data augmentation and a appropriate combination of CNNs, results that are competitive against the state-of-the-art methods can be achieved. To make a

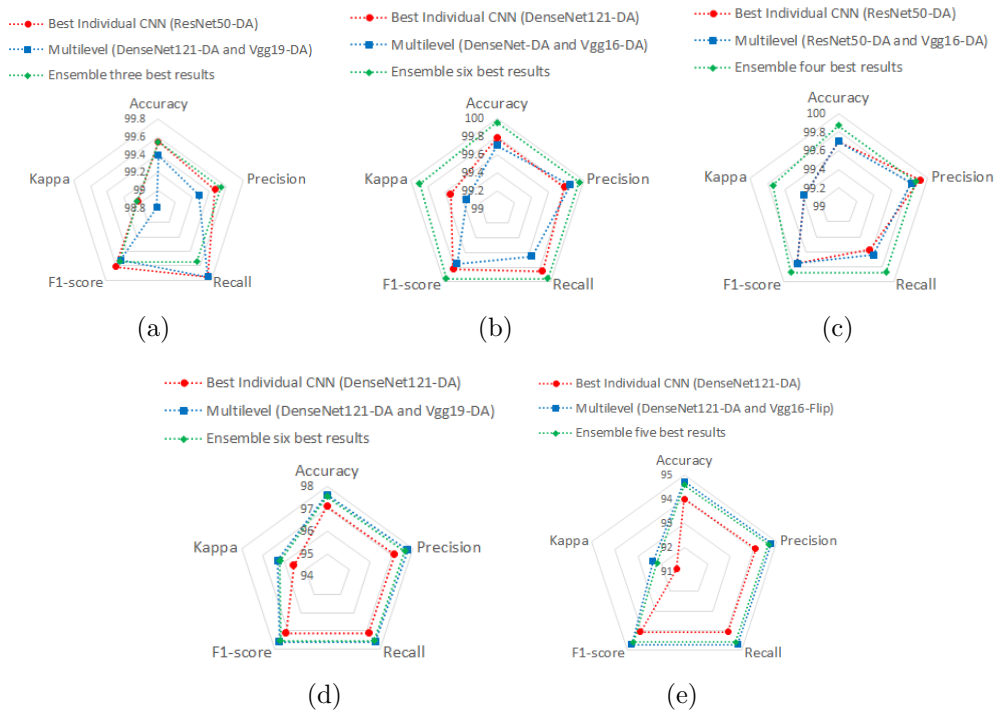


Figure 7: Comparison of the best results obtained by the individual, multilevel and ensemble CNN configurations in the (a) leukemia vs. HBS, (b) ALL vs. HBS, (c) AML vs. HBS, (d) ALL vs. AML vs. HBS and (e) ALL vs. AML vs. HBS vs. other types, scenarios.

more reliable comparison, Table 11 is organized according to the type of the addressed classification.

In distinguishing leukemia from HBS, excellent results were achieved by the work of Loey et al. (2020), which reached 100% of accuracy. It can be noted that the ensemble and multilevel configurations used in this study obtained excellent accuracies. However, although used was used the AlexNet network in the current study, as suggested by Loey et al. (2020), it did not perform well, which could be due to the fact that its depth is much smaller than the depths of the other studied networks. Therefore, it was difficult for this network to learn the features of the used image datasets.

The ensemble and multilevel configurations achieved better performance in classifying images for the ALL and HBS and for AML and HBS scenarios. The ensemble configuration obtained 99.95% of accuracy for the first scenario,

and 99.87% of accuracy for the second one. One could see that the other works that used these classifiers did not used images of high diversity. In fact, most of them used only one dataset, which does not lead to a robust classifier.

As to the classification into three classes: ALL, AML and HBS, the study by Laosai and Chamnongthai (2018) obtained the best results found, with 99.85% of accuracy on 500 images from two private datasets. In our test using a dataset of 3,293 publicly available images, which is almost seven times as large as the dataset used by Loey et al. (2020), the multilevel and ensemble configurations achieved accuracies of 97.60 and 97.53%, respectively.

One can see that most of the state-of-the-art works presented higher accuracy values than those of the proposed models. However, the number of images used in those studies is lower than the one used here. Another detail to be highlighted is that most related articles used data augmentation techniques to reduce overfitting. It is also important to emphasize that a heterogeneous set of images from eighteen publicly available datasets with different image characteristics, such as illumination, contrast and brightness, was used in the current study, which led to a greater diversity in the training data, allowing one to obtain a more robust classifier for different input images.

To the best of our knowledge, the division into ALL, AML, HBS, and other types of leukemia addressed here was the first study to explore all four classes of leukemia types. According to Table 11, it was obtained 94.73% of accuracy using the multilevel configuration and 94.59% of accuracy using the ensemble configuration.

Figure 8 presents examples of the activation maps obtained by the VGG16 and DenseNet121 networks for the scenario with four classes. In this figure, the first column presents the original images, and the second column corresponds to the results obtained using the VGG16 network. For this network, the blue shades indicate a low activation which suggests that the corresponding regions are of low importance for the final classification; conversely, red tones are associated with the areas that are the most critical to the final classification. In the same figure, the third column presents the results obtained using DenseNet121, and unlike the VGG16 network, the parts in blue are those associated with the most critical regions for the final classification. These regions of the peripheral blood smear are of great importance to clinicians examining the appearance of cells, as changes in cell numbers and appearance often help to diagnose leukemia.

Table 11: Comparison among the results obtained using the proposed multilevel and ensemble configurations and the results obtained using related methods.

Method	Classification	Number of Datasets	Number of images	Accuracy(%)
Thanh et al. (2018)	Leukemia - HBS	1	108	96.60
Vogado et al. (2018)	Leukemia - HBS	8	1,268	99.76
Loey et al. (2020)	Leukemia - HBS	2	564	100
Vogado et al. (2021)	Leukemia - HBS	18	3,536	98.61
Multilevel	Leukemia - HBS	18	3,536	99.39
Ensemble	Leukemia - HBS	18	3,536	99.54
Singhal and Singh (2016)	ALL - HBS	1	260	93.80
Shafique and Tehsin (2018)	ALL - HBS	2	368	99.50
Ahmed et al. (2019)	ALL - HBS	2	354	88.25
Pansombut et al. (2019)	ALL - HBS	2	363	81.74
Gehlot et al. (2020)	ALL - HBS	1	15,114	93.40
Zakir Ullah et al. (2021)	ALL - HBS	1	15,114	91.10
Karar et al. (2022)	ALL - HBS	1	368	98.65
Rodrigues et al. (2022)	ALL - HBS	1	260	98.46
Abhishek et al. (2022)	ALL - HBS	2	608	97.00
Multilevel	ALL - HBS	14	2,315	99.70
Ensemble	ALL - HBS	14	2,315	99.95
Madhukar et al. (2012)	AML - HBS	1	50	93.50
Goutam and Sailaja (2015)	AML - HBS	1	90	98.00
Dasariraju et al. (2020)	AML - HBS	1	1,274	92.99
Multilevel	AML - HBS	16	2,412	99.67
Ensemble	AML - HBS	16	2,412	99.87
Rawat et al. (2017)	ALL - AML - HBS	1	240	99.50
Tran et al. (2018)	ALL - AML - HBS	2	141	97.30
Laosai and Chamnongthai (2018)	ALL - AML - HBS	2	500	99.85
Claro et al. (2020)	ALL - AML - HBS	16	2,415	97.18
Karar et al. (2022)	ALL - AML - HBS	2	445	95.50
Abhishek et al. (2022)	ALL - AML - HBS	2	608	98.00
Multilevel	ALL - AML - HBS	18	3,293	97.60
Ensemble	ALL - AML - HBS	18	3,293	97.53
Multilevel	ALL - AML - HBS - other Types	18	3,536	94.73
Ensemble	ALL - AML - HBS - other Types	18	3,536	94.59

An alternative that was not evaluated in this study is enhancing the input images before starting the classification process. For example, a proper contrast transformation can improve the perceptual quality of the most used images (Gu et al., 2015). In fact, according to Gu et al. (2016), a good image enhancement transformation can noticeably improve the quality of the input images, so that they are even better than the originally acquired images, which are generally thought to be of high quality. However, the choice of the image enhancement transform should be prudent, since most relevant technologies often suffer from the drawback of excessive enhancement, thus introducing noise/artifacts and changing the visual attention regions (Gu et al., 2018).

Another essential point to be evaluated is using CNNs that are pre-trained on ImageNet. According to Kornblith et al. (2019), the architec-tures that perform best on ImageNet can provide better feature extraction and fine-tuning. However, the authors observed this fact only in photo-graphic datasets. According to Sipes and Li (2018), leukemia images are considered fine-grained images. Classification tasks using datasets with fine-

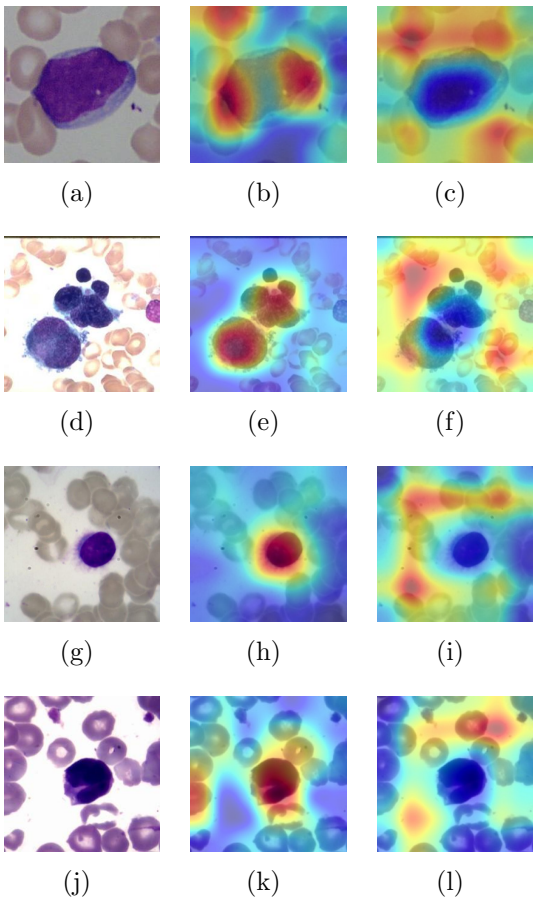


Figure 8: Examples of activation maps obtained for the imaging blood slides under study: (a-c) ALL images, (d-f) AML images, (g-i) chronic leukemia images and (j-l) HBS images.

grained images can be particularly challenging, and the effects of pre-training with ImageNet were deemed to be small. Therefore, it is essential to evaluate the use of specific refinement techniques for fine-grained images (Du et al., 2021).

In the tests conducted in this study, the original CNN architectures were used. However, changes in these architectures can lead to performance improvements. For example, Claro et al. (2020) performed an ablation procedure to define a specific acute leukemia CNN. Another possibility is the insertion of attention mechanisms, such as the one proposed by Xie et al. (2021). The core idea of the attention mechanisms is that each time the

model predicts an output, it only uses the parts of the input where the most relevant information is concentrated.

This study assessed the multilevel and ensemble combinations, but other configurations exist for CNN-based models. For example, hierarchical models can be used for problems with several classes. These models exploit the hierarchical structure of object categories to decompose classification tasks into multiple steps. They have shown better performance than flat models in performing image classification across multiple domains (Kowsari et al., 2020).

6. Conclusion

In this study, techniques that can be integrated into computer-aided diagnostic systems in order to detect different types of leukemia, mainly ALL, AML, and other types, in addition to healthy slides, were evaluated. Several experiments were We performed. First, tests were performed according to five scenarios and the effectiveness of using techniques of data augmentation was analyzed. Then, a comparison among techniques of data augmentation for the ALL vs. AML vs. HBS vs. other types classification was performed. The total number of the used images, which were gathered from different public datasets, was equal to 3,536. Data augmentation is becoming an im-portant area of research, particularly in medical imaging. There are articles on techniques of data augmentation for specific types of images. However, there is not enough comprehensive work on leukemia images augmentation and classification. Therefore, this article will also benefit many researchers working on this area.

The concept of multilevel and ensemble configurations was applied in the conducted experiments, aiming to increase the accuracy and generalization of the classification and decrease its standard deviation rate. It is essential to highlight the use of the ensemble configuration, because it reduces the standard deviation rate of the classification by optimally combining the predictions of several models. The built ensemble model simulates real-world conditions leading to reduced standard deviation and overfitting, and enhanced generalization. One can concluded that the achieved findings can aid the development of clinically valuable solutions for detecting and differentiating among different leukemia types in blood slide images.

In near future, more tests will be performed using an even larger number of testing images, and an image enhancement step will be included into the

proposed classification configuration. Future research will also investigate the differences between cell images with chronic lymphoid leukemia and chronic myeloid leukemia. Additionally, a literature survey will be conducted to identify other new insights concerning these types of leukemia.

Acknowledgment

This study was partially founded by the "Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior" (CAPES) - Finance Code 001, "Fundação de Amparo a Pesquisa do Piaui "(FAPEPI), and "Conselho Nacional de Desenvolvimento Cientifico e Tecnologico" (CNPQ), in Brazil. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used in this study.

References

Abbas, N., Mohamad, D., 2014. Automatic color nuclei segmentation of leukocytes for acute leukemia. Research Journal of Applied Sciences, Engineering and Technology 7, 2987–2993. doi:10.19026/rjaset.7.631.

Abhishek, A., Jha, R.K., Sinha, R., Jha, K., 2022. Automated classification of acute leukemia on a heterogeneous dataset using machine learning and deep learning techniques. Biomedical Signal Processing and Control 72, 103341.

Ahmed, N., Yigit, A., Isik, Z., Alpkocak, A., 2019. Identification of leukemia subtypes from microscopic images using convolutional neural network. Diagnostics 9, 104. doi:10.3390/diagnostics9030104.

Anilkumar, K., Manoj, V., Sagi, T., 2020. A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of leukemia. Biocybernetics and Biomedical Engineering 40, 1406–1420. doi:10.1016/j.bbe.2020.08.010.

Arslan, S., Ozyurek, E., Gunduz-Demir, C., 2014. A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images. Cytometry Part A 85, 480–490. doi:10.1002/cyto.a.22457.

ASH, 2020. Ash image bank: American society of hematology. URL: http://imagebank.hematology.org/Default.aspx.

Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences 116, 15849–15854.

Bibi, N., Sikandar, M., Ud Din, I., Almogren, A., Ali, S., 2020. Iomt-based automated detection and classification of leukemia using deep learning. Journal of Healthcare Engineering 1, 1–12. doi:10.1155/2020/6648574.

Böhm, J., 2008. Pathologie-websites im world wide web. Der Pathologe 29, 231–242.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 .

Chen, X., 2019. Image enhancement effect on the performance of convolutional neural networks.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Claro, M., Vogado, L., Veras, R., Santana, A., Tavares, J., Santos, J., Machado, V., 2020. Convolution neural network models for acute leukemia diagnosis, in: International Conference on Systems, Signals and Image Processing, IEEE. pp. 63–68. doi:10.1109/IWSSIP48289.2020.9145406.

Das, P.K., Meher, S., 2021. An efficient deep convolutional neural network based detection and classification of acute lymphoblastic leukemia. Expert Systems with Applications 183, 115311. doi:10.1016/j.eswa.2021.115311.

Dasarathy, B.V., Sheela, B.V., 1979. A composite classifier system design: Concepts and methodology. Proceedings of the IEEE 67, 708–713. doi:10.1109/PROC.1979.11321.

Dasariraju, S., Huo, M., McCalla, S., 2020. Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm. Bioengineering 7, 120.

Dietterich, T.G., 2000. Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer. pp. 1–15. doi:https://doi.org/10.1007/3-540-45014-9_1.

Du, R., Xie, J., Ma, Z., Chang, D., Song, Y.Z., Guo, J., 2021. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/TPAMI.2021.3126668.

Gehlot, S., Gupta, A., Gupta, R., 2020. Sdct-auxnetθ: Dct augmented stain deconvolutional cnn with auxiliary classifier for cancer diagnosis. Medical Image Analysis 61, 101661. doi:https://doi.org/10.1016/j.media.2020.101661.

Goceri, E., 2020. Image augmentation for deep learning based lesion classification from skin images, in: Proceedings of the IEEE 4th International Conference on Image Processing, Applications and Systems, pp. 144–148. doi:10.1109/IPAS50080.2020.9334937.

Goutam, D., Sailaja, S., 2015. Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier, in: Proceedings of the IEEE International Conference on Engineering and Technology, IEEE. pp. 1–5. doi:10.1109/ICETECH.2015.7275021.

Gu, K., Tao, D., Qiao, J.F., Lin, W., 2018. Learning a no-reference quality assessment model of enhanced images with big data. IEEE Transactions on Neural Networks and Learning Systems 29, 1301–1313. doi:10.1109/TNNLS.2017.2649101.

Gu, K., Zhai, G., Lin, W., Liu, M., 2016. The analysis of image contrast: From quality assessment to automatic enhancement. IEEE Transactions on Cybernetics 46, 284–297. doi:10.1109/TCYB.2015.2401732.

Gu, K., Zhai, G., Yang, X., Zhang, W., Chen, C.W., 2015. Automatic contrast enhancement technology with saliency preservation. IEEE Transactions on Circuits and Systems for Video Technology 25, 1480–1494. doi:10.1109/TCSVT.2014.2372392.

Gupta, A., Gupta, R., Gehlot, S., Mourya, S., 2019. Classification of normal vs malignant cells in b-all white blood cancer microscopic images, in:

IEEE International Symposium on Biomedical Imaging (ISBI) Challenges Internet, pp. 1–12. doi:10.1007/978-981-15-0798-4_1.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 770–778. doi:10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708. doi:10.1109/CVPR.2017.243.

Janocha, K., Czarnecki, W.M., 2017. On loss functions for deep neural networks in classification. arXiv:1702.05659 .

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM international conference on Multimedia, pp. 675–678. doi:10.1145/2647868.2654889.

Jiang, K., Liao, Q.M., Dai, S.Y., 2003. A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering, in: Proceedings of the International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693), pp. 2820–2825. doi:10.1109/ICMLC.2003.1260033.

Karar, M.E., Alotaibi, B., Alotaibi, M., 2022. Intelligent medical iot-enabled automated microscopic image diagnosis of acute blood cancers. Sensors 22, 2348.

Khan, S., Sajjad, M., Hussain, T., Ullah, A., Imran, A.S., 2020. A review on traditional machine learning and deep learning models for wbcs classification in blood smear images. IEEE Access 9, 10657–10673. doi:10.1109/ACCESS.2020.3048172.

Khandekar, R., Shastry, P., Jaishankar, S., Faust, O., Sampathila, N., 2021. Automated blast cell detection for acute lymphoblastic leukemia diagnosis. Biomedical Signal Processing and Control 68, 102690. doi:https://doi.org/10.1016/j.bspc.2021.102690.

Khosla, E., Ramesh, D., 2018. Phase classification of chronic myeloid leukemia using convolution neural networks, in: Proceedings of the 4th International Conference on Recent Advances in Information Technology, IEEE. pp. 1–6. doi:10.1109/RAIT.2018.8389068.

Kim, Y.j., Cho, H.C., Cho, H.c., 2021. Deep learning-based computer-aided diagnosis system for gastroscopy image classification using synthetic data. Applied Sciences 11, 760. doi:10.3390/app11020760.

Kornblith, S., Shlens, J., Le, Q.V., 2019. Do better imagenet models transfer better?, in: Proceedings if the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2661–2671.

Kowsari, K., Sali, R., Ehsan, L., Adorno, W., Ali, A., Moore, S., Amadi, B., Kelly, P., Syed, S., Brown, D., 2020. Hmic: Hierarchical medical image classification, a deep learning approach. Information 11, 318. URL: https://www.mdpi.com/2078-2489/11/6/318, doi:10.3390/info11060318.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105.

Kukačka, J., Golkov, V., Cremers, D., 2017. Regularization for deep learning: A taxonomy. arXiv preprint arXiv:1710.10686 .

Labati, R.D., Piuri, V., Scotti, F., 2011. All-idb: The acute lymphoblastic leukemia image database for image processing., in: Proceedings of the IEEE International Conference on Image Processing, pp. 2045–2048. doi:10.1109/ICIP.2011.6115881.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 3, 159–174. doi:10.2307/2529310.

Laosai, J., Chamnongthai, K., 2018. Classification of acute leukemia using medical-knowledge-based morphology and cd marker. Biomedical Signal Processing and Control 44, 127–137. doi:https://doi.org/10.1016/j.bspc.2018.01.020.

Li, L., Liang, J., Weng, M., Zhu, H., 2018. A multiple-feature reuse network to extract buildings from remote sensing imagery. Remote Sensing 10, 1350. doi:10.3390/rs10091350.

Li, S., Gong, K., Liu, C.H., Wang, Y., Qiao, F., Cheng, X., 2021. Metasaug: Meta semantic augmentation for long-tailed visual recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5212 – 5221.

Li, Y., Zhu, R., Mi, L., Cao, Y., Yao, D., 2016. Segmentation of white blood cell from acute lymphoblastic leukemia images using dual-threshold method. Computational and mathematical methods in medicine 2016, 1–12. doi:10.1155/2016/9514707.

Loey, M., Naman, M., Zayed, H., 2020. Deep transfer learning in diagnosing leukemia in blood cells. Computers 9, 29. doi:10.3390/computers9020029.

Lyu, J., Ling, S.H., 2018. Using multi-level convolutional neural network for classification of lung nodules on ct images, in: Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 686–689. doi:10.1109/EMBC.2018.8512376.

Madhukar, M., Agaian, S., Chronopoulos, A.T., 2012. Deterministic model for acute myelogenous leukemia classification, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE. pp. 433–438. doi:10.1109/ICSMC.2012.6377762.

Mishra, S., Majhi, B., Sa, P.K., 2016. A survey on automated diagnosis on the detection of leukemia: A hematological disorder, in: Proceedings of the 3rd International Conference on Recent Advances in Information Technology, IEEE. pp. 460–466. doi:10.1109/RAIT.2016.7507945.

Mrozek, K., Heerema, N.A., Bloomfield, C.D., 2004. Cytogenetics in acute leukemia. Blood reviews 18, 115–136. doi:https://doi.org/10.1016/S0268-960X(03)00040-7.

Pansombut, T., Wikaisuksakul, S., Khongkraphan, K., Phon-on, A., 2019. Convolutional neural networks for recognition of lymphoblast cell images. Computational Intelligence and Neuroscience 2019, 1–12. doi:10.1155/2019/7519603.

Powers, D.M., 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, School of Informatics and

Engineering, Flinders University, Adelaide, Australia. Technical Report 1. TR SIE-07-001, Journal of Machine Learning Technologies.

Puttagunta, M., Ravi, S., 2021. Medical image analysis based on deep learning approach. Multimedia Tools and Applications 80, 24365–24398. doi:10.1007/s11042-021-10707-4.

Rastogi, P., Khanna, K., Singh, V., 2022. Leufeatx: Deep learning–based feature extractor for the diagnosis of acute leukemia from microscopic images of peripheral blood smear. Computers in Biology and Medicine , 105236.

Rawat, J., Singh, A., HS, B., Virmani, J., Devgun, J.S., 2017. Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia. Biocybernetics and Biomedical Engineering 37, 637 – 654. doi:10.1016/j.bbe.2017.07.003.

Rezatofighi, S.H., Soltanian-Zadeh, H., 2011. Automatic recognition of five types of white blood cells in peripheral blood. Computerized Medical Imaging and Graphics 35, 333 – 343. doi:https://doi.org/10.1016/j.compmedimag.2011.01.003.

Rodrigues, L.F., Backes, A.R., Travençolo, B.A.N., de Oliveira, G.M.B., 2022. Optimizing a deep residual neural network with genetic algorithm for acute lymphoblastic leukemia classification. Journal of Digital Imaging , 1–15.

Rollins-Raval, M., Raval, J., Contis, L., 2012. Experience with cellavision dm96 for peripheral blood differentials in a large multi-center academic hospital system. Journal of Pathology Informatics 3, 1–9. doi:10.4103/2153-3539.100154.

Sarrafzadeh, O., Dehnavi, A.M., 2015. Nucleus and cytoplasm segmentation in microscopic images using k means clustering and region growing. Advanced Biomedical Research 4, 79–87. doi:10.4103/2277-9175.163998.

Sarrafzadeh, O., Rabbani, H., Dehnavi, A.M., Talebi, A., 2015. Detecting different sub-types of acute myelogenous leukemia using dictionary learning and sparse representation, in: Proceedings of the IEEE International Conference on Image Processing, IEEE. pp. 3339–3343. doi:10.1109/ICIP.2015.7351422.

Sarrafzadeh, O., Rabbani, H., Talebi, A., Banaem, H.U., 2014. Selection of the best features for leukocytes classification in blood smear microscopic images, in: SPIE Medical Imaging, p. 90410P. doi:10.1117/12.2043605.

Sarvamangala, D.R., Kulkarni, R.V., 2021. Convolutional neural networks in medical image understanding: a survey. Evolutionary Intelligence , 1 – 22doi:10.1007/s12065-020-00540-3.

Shafique, S., Tehsin, S., 2018. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technology in Cancer Research and Treatment 17, 1–7. doi:10.1177/1533033818802789.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6, 60. doi:10.1186/s40537-019-0197-0.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Singhal, V., Singh, P., 2016. Texture Features for the Detection of Acute Lymphoblastic Leukemia. Springer Singapore, Singapore. pp. 535–543. doi:10.1007/978-981-10-0135-2_52.

Sipes, R., Li, D., 2018. Using convolutional neural networks for automated fine grained image classification of acute lymphoblastic leukemia, in: Proceedings of the 3rd International Conference on Computational Intelligence and Applications, pp. 157–161. doi:10.1109/ICCIA.2018.00036.

Society, A.C., 2021. Leukemia statistics. URL: https://cancerstatisticscenter.cancer.org/#!/cancer-site/Leukemia.

Sollich, P., Krogh, A., 1995. Learning with ensembles: How overfitting can be useful. Advances in neural information processing systems 8, 190–196.

Souza, L.M.d., Gorini, M.I.P.C., 2006. Diagnósticos de enfermagem em adultos com leucemia mielóide aguda. Revista gaúcha de enfermagem. Porto Alegre. 27, 417–425.

Srinivas, M., Bharath, R., Rajalakshmi, P., Mohan, C.K., 2015. Multi-level classification: A generic classification method for medical datasets, in:

Proceedings of the 17th International Conference on E-health Networking, Application & Services (HealthCom), IEEE. pp. 262–267. doi:10.1109/HealthCom.2015.7454509.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826. doi:10.1109/CVPR.2016.308.

Thanh, T., Vununu, C., Atoev, S., Lee, S.H., Kwon, K.R., 2018. Leukemia blood cell image classification using convolutional neural network. International Journal of Computer Theory and Engineering 10, 54–58. doi:10.7763/IJCTE.2018.V10.1198.

Tran, T., Park, J.H., Kwon, O.H., Moon, K.S., Lee, S.H., Kwon, K.R., 2018. Classification of leukemia disease in peripheral blood cell images using convolutional neural network. Journal of Korea Multimedia Society 21, 1150–1161. doi:10.9717/kmms.2018.21.10.1150.

Travlos, G.S., 2006. Normal structure, function, and histology of the bone marrow. Toxicologic Pathology 34, 548–565. doi:10.1080/01926230600939856.

Upreti, M., Pandey, C., Bist, A.S., Rawat, B., Hardini, M., 2021. Convolutional neural networks in medical image understanding. Aptisi Transactions on Technopreneurship 3, 6–12. doi:10.1007/s12065-020-00540-3.

Vale, A.M.P.G., Guerreiro, A.M.G., Neto, A.D.D., Cavalvanti Junior, G.B., de Sá Leitão, V.C.L.T., Martins, A.M., 2014. Automatic segmentation and classification of blood components in microscopic images using a fuzzy approach. Revista Brasileira de Engenharia Biomédica 30, 341–354. doi:10.1590/1517-3151.0626.

Vogado, L., Veras, R., Aires, K., Araújo, F., Silva, R., Ponti, M., Tavares, J.M.R., 2021. Diagnosis of leukaemia in blood slides based on a fine-

tuned and highly generalisable deep learning model. Sensors 21, 2989. doi:10.3390/s21092989.

Vogado, L.H., Veras, R.d.M., Andrade, A.R., e Silva, R.R., De Araujo, F.H., De Medeiros, F.N., 2016. Unsupervised leukemia cells segmentation based on multi-space color channels, in: Proceedings of the IEEE International Symposium on Multimedia, IEEE. pp. 451–456. doi:10.1109/ISM.2016.0103.

Vogado, L.H.S., Veras, R.M.S., Araújo, F.H.D., e Silva, R.R.V., Aires, K.R.T., 2018. Leukemia diagnosis in blood slides using transfer learning in cnns and SVM for classification. Engineering Applications of Artificial Intelligence 72, 415–422. doi:10.1016/j.engappai.2018.04.024.

Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., Wu, C., 2021. Regularizing deep networks with semantic data augmentation. arXiv:2007.10538.

Xie, J., Ma, Z., Chang, D., Zhang, G., Guo, J., 2021. Gpca: A probabilistic framework for gaussian process embedded channel attention. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/TPAMI.2021.3102955.

Zakir Ullah, M., Zheng, Y., Song, J., Aslam, S., Xu, C., Kiazolu, G.D., Wang, L., 2021. An attention-based convolutional neural network for acute lymphoblastic leukemia classification. Applied Sciences 11, 10662. doi:10.3390/app112210662.

Zheng, X., Wang, Y., Wang, G., Chen, Z., 2018. Fast and robust segmentation of white blood cell images by self-supervised learning. Micron 107, 55–71. doi:https://doi.org/10.1016/j.micron.2018.01.010.