Detecting Acute Leukemia in Blood Slides Images Using a CNNs Ensemble

Abstract: Leukemia is a disease that has no defined etiology and affects the production of white blood cells in the bone marrow. Young cells or blasts are produced abnormally, replacing normal blood cells (white, red blood cells, and platelets). Consequently, the person suffers problems in transporting oxygen and infections combat. Acute leukemia is a particular type of leukemia that causes abnormal cell growth in a short period, requiring a quick start of treatment. Classifying the types of acute leukemia in blood slide images is a vital process, and a system of assisting doctors in selecting treatment becomes necessary. This paper presents an ensemble approach using four convolutional neural networks (CNNs) - Alert Net-RWD, Resnet50, InceptionV3, and Xception. These CNNs, individually, demonstrated that are efficient in differentiating between the two types of acute leukemia - Acute Lymphoid Leukemia (ALL) and Acute Myeloid Leukemia (AML) - and Healthy Blood Slides (HBS). We verified that the union of these four well-known CNNs improve the hit rates of current techniques from the literature. The experiments were carried out using 18 data sets with 3,293 images, and the proposed CNNs ensemble achieved an accuracy of 96.17%, and precision of 96.38%.

Keywords: Acute leukemia diagnosis, model ensemble, convolutional neural network.

1 Introduction

The bone marrow produces a large proportion of blood cells, 100 million of leucocytes (white blood cells) per day on average. Leukocytes act combating and eliminating microorganisms and foreign chemical structures in the body employing a catch (phagocytosis) or antibody production. One of the diseases affecting the bone marrow function is leukemia [1].

Leukemia is a type of cancer that mostly affects the population. The American Cancer Society (ACS) (https://cancerstatisticscenter.cancer.org/!/cancer-site/Leukemia) estimates 60,530 new cases for 2020, with approximately 23,100 deaths. According to the ACS, 35,470 cases are in men and 25,060 in women and 13,420 deaths in men and 9,860 in women. This disease has no defined etiology and affects the production of cells by the bone marrow. Over time, diseased cells replace healthy blood cells (white, red blood cells, and platelets), and the individual suffers from problems in transporting oxygen and fighting infections [1]. Among the forms of leukemia diagnosis, the complete blood count (CBC) and the myelogram are the most used.

Leukemias can be grouped based on the speed at which the disease progresses and becomes severe. In this respect, the condition can be of the chronic type (which usually gets worse slowly) or acute (which usually gets worse quickly) [1]. They can also be grouped based on the types of white blood cells they affect: lymphoid or myeloid. Thus, there are some types of leukemia, the four primary ones being Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML) and Chronic Lymphocytic Leukemia (CLL). Myeloid leukemia (SCI) is the most frequent, with about 40% of diagnoses.

Thus, ALL and AML require a diagnosis in the early stages of the disease to provide appropriate treatment. Figure 1 shows examples of blood slide images used in our tests with ALL, AML, and Healthy Blood Slides (HBS). The first column of Figure 1 shows examples of images with ALL; the second column shows images with AML, and the last column shows HBS images.

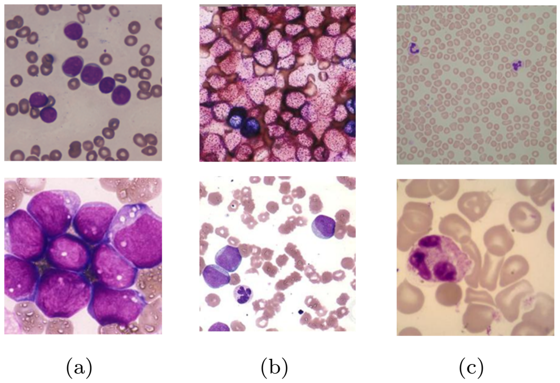


Figure 1 Examples of images used in this work: (a) ALL example, (b) AML example, and (c) HBS example.

The use of computer systems can assist in fast leukemia diagnosis. Convolutional Neural Networks (CNNs) is currently one of the most effective techniques in diagnosing medical images. This paper proposes an ensemble model with CNNs for the automated diagnosis of acute leukemia (ALL and AML) and HBS. Resnet50 [2], inceptionV3 [3], Xception [4] and the network models proposed by Claro et al. [5]. Also, we implemented a new network based on the last-mentioned work. We evaluated the proposed model in 18 heterogeneous datasets with 3,293 images, combined with data augmentation techniques.

This paper is organized as follows. Section 2 presents related works; In Section 3, we present the dataset used, the data augmentation technique employed, the evaluated architectures and the applied evaluation metrics. Sections 4 and 5 present the achieved results and a discussion; and finally, we present the conclusion and possibilities of future work in Section 6.

2 Related Works

We carried out a systematic survey of the state of the art related to leukemia computer-aided diagnosis. The survey aimed to identify and classify the available literature works based on the techniques used, the year of publication, and the relevance.

The survey was realized using three public datasets: Scopus, Web of Science, and IEEE Xplore. We used the following search strings: "leukemia acute classification", "white blood cell classification", and "blood smear leukemia classification". Following this, we selected works published after 2012 in engineering and computer science fields. As a result, we obtained 427 articles. We then analyzed the title and abstract of these, aiming to eliminate repeated documents and those with non-automatic classification methods. Table 1 presents the works found in the literature, organized according to their purpose.

We organized the selected papers into four approaches using the diagnosis type suggested by the authors. We found studies that performed the diagnosis between images with leukemia and healthy, regardless of the type of leukemia [6] and [7]. Some authors differentiated blades of blood with ALL and healthy blades [8, 9, 10, 11, 12, 13, 14, 15], while other proposals differentiated images with AML and healthy images [16, 17].

The latter approach mentioned above coincides with our proposal and is characterized by the diagnosis into three classes: ALL, AML, and HBS. In Rawat et al. [18], the authors performed the leukocyte nucleus segmentation on 420 images. Then, they analyzed 331 characteristics of each segmented nucleus using a Support Vector Machine (SVM). The work of Laosai and Chamnongthai [19] also to take into account these categories. They subdivide the types of acute lymphoid leukemia and acute myeloid leukemia. The tests were performed on 500 images, 150 of ALL type, 150 of AML type, and 200 of HBS type. According to the authors, the tests showed promising results.

Still, in this approach, we have the work of Tran et al. [20] and Claro et al. [5]. In both studies, the authors proposed convolutional neural networks to classify the two types of acute leukemia and images without leukemia. In the first work, the system developed was LeukemiaNet, and in the second work, the network presented was Alert Net-RWD.

3 Materials and Methods

This paper aims to present models of CNN architectures to diagnose acute leukemia types in blood slide images. To develop the architectural model proposed in this work, we rely on architectures that recently obtained the best results in leukemia detection, according to the studies found in the literature.

The dataset used in this research hold 3,293 images, which does not represent a large number of data for training a CNN. A solution found to increase the generality of the model and attack the few cases problem for training the network is the Data Augmentation technique that generates new training samples.

3.1 Image Dataset

The development of a robust methodology to aid in the diagnosis depends on the data used in its validation. The main challenge found in state of the art is related to the datasets' acquisition since most of them are private. However, we obtained 18 public datasets with 3,293 images for the evaluation of the proposed model. In Table 2, the used image datasets are presented according to the addressed classes.

Among the images listed in Table 2, we disregarded those "Other Types" class, since the amount of data in this class does not form an adequately representative set. Thus, we used the HBS, ALL, and AML classes to build the proposed model. One can note that these classes were made using different datasets, contributing to the creation of a complex set with different resolutions, dyes, approximations, and contrast. The before-mentioned approach is similar to the one used to obtain microscopic images in daily medical practices [14].

Figure 2 shows examples of blood slide images used in our tests with ALL, AML, and Healthy Blood Slides (HBS). These images are some examples of the databases used in this study.

For the proper use of the images under study, we carried out two pre-processing operations. The first was the central clipping considering the smaller side image since CNN architectures require square inputs. The second operation was to resize the input images to 224 × 224 pixels because these are the standard CNN input dimensions.

In the Bloodline dataset, we observed the existence of 15 rectangular images containing at least two leukocytes. In these images, the leukocytes' clipping was done manually, since the pre-processing operations would eliminate the region of interest for the classification. The clipped images were added to the dataset, resulting in a total of 217 samples.

3.2 Data Augmentation

Deep neural networks have been successfully applied to Computer Vision tasks such as image classification, object detection, and image segmentation, thanks to the

Table 1 Summary of works identified in the state of the art as to: year, descriptor(s), classifier, number of images used and accuracy.

Work	Year	Descriptor(s)	Classifier	Images	Accuracy(%)
Leukemia - Healthy					
Vogado et al. [6]	2018	*Deep Features*	SVM	1,268	99.76
Loey et al. [7]	2020	*Deep Features*	CNN	2,820	100
ALL - Healthy					
Vincent et al. [8]	2015	Texture and geometry	MLP	100	97.70
Singhal and Singh [9]	2016	Texture	SVM	260	93.80
Thanh et al. [10]	2017	*Deep Features*	CNN	1,188	96,60
Shafique et al. [11]	2018	*Deep Features*	CNN	760	99.50
Rehman et al. [12]	2018	*Deep Features*	CNN	330	97.78
Pansombut et al. [13]	2019	*Deep Features*	CNN	363	81.74
Ahmed et al. [14]	2019	*Deep Features*	CNN	2,478	88.25
Gehlot et al. [15]	2020	*Deep Features*	CNN	15,114	93.40
AML - Healthy					
Madhukar and Agaian [16]	2012	Texture	SVM	50	93.50
Goutam and Sailaja [17]	2015	Texture	SVM	90	98.00
ALL - AML - Healthy					
Rawat et al. [18]	2017	Geometrical, color and texture	GA-SVM	420	99.50
Laosai and Chamnongthai [19]	2018	Shape, color distribution, texture and number of nucleoli	SVM	500	99.85
Tran et al. [20]	2018	Deep Features	CNN	1,676	97.20
Claro et al. [5]	2020	Deep Features	CNN	2,415	97.18

Table 2 Summary of the used image datasets.

Dataset	HBS	ALL	AML	Other types	Total	Ref.
ALL-IDB 1	59	49	-	-	108	[21]
ALL-IDB 1 (Crop)	-	510	-	-	510	[21]
ALL-IDB 2	130	130	-	-	260	[21]
Leukocytes	149	-	-	-	149	[22]
CellaVision	109	-	-	-	109	[23]
Atlas	-	25	40	23	88	-
Omid et al. 2014	154	-	-	-	154	[24]
Omid et al. 2015	-	-	27	-	27	[25]
ASH-OK	-	-	96	-	96	[26]
Bloodline	-	-	217	-	217	[27]
ONKODIN	-	-	78	-	78	[28]
CellaVision 2	100	-	-	-	100	[29]
JTSC	300	-	-	-	300	[29]
UFG	57	10	27	27	121	-
PN-ALL Dataset	-	30	-	-	30	[30]
leukemia-images	-	40	78	22	140	-
MIDB Dataset	-	87	415	171	673	-
LISC Dataset	376	-	-	-	376	[31]
Total of images	**1434**	**881**	**978**	**243**	**3536**	-

evolution of CNNs. However, these networks rely on a large amount of data to avoid overfitting [32].

Improving generalization of these models is one of the main challenges in the area, but Data Augmentation is a powerful way to overcome this difficulty. Augmented data is expected to represent a more extensive dataset, minimizing the differences between the training and validation sets as well as any future test sets [32].

Routine augmentation operations are rotation in the range of 0º to 40º, vertical, horizontal, shear, and zoom in the field of 0 to 0.2 and horizontal and vertical flip. One should notice that the nuclei images do not have asymmetry allowing flipping in both directions. The reflection fill operation was applied to replace black pixels resulting from rotation and translation techniques. Finally, we normalized the input image pixels to values between 0 (zero) and 1 (one). The augmentation resulted in a dataset 20 times bigger than the original. Figure 3 shows examples of the results of these operations applied in blood smear images.

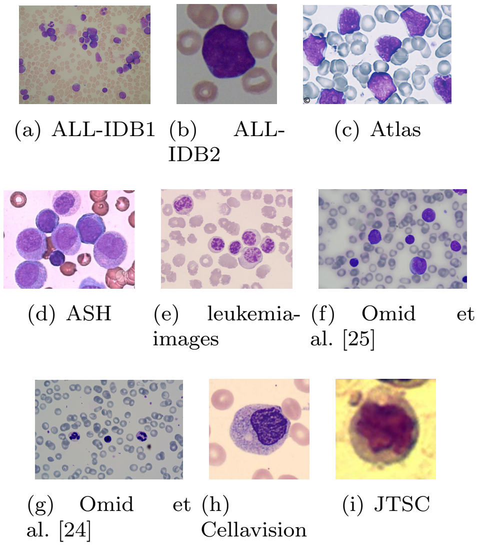

(a) ALL-IDB1 (b) ALL-IDB2 (c) Atlas
(d) ASH (e) leukemia-images (f) Omid et al. [25]
(g) Omid et al. [24] (h) Cellavision (i) JTSC

Figure 2 Samples of (a-c) ALL images, (d-f) AML images and (g-i) HBS images used in this work.

3.3 Evaluated Architectures

In this study, we evaluated the pre-trained neural networks Resnet50 [2], InceptionV3 [3] and Xception [4] to classify the two types of acute leukemia (ALL and AML) and healthy images. In addition to these architectures, we also evaluated the models developed by Claro et al. [5] and a new variation developed in this work based on the proposed models.

ResNet50 [2] is a deep convolutional network architecture proposed in 2016 to solve the problem of *vanishing gradient* that causes saturation in learning and consequently slows down training. The basic idea is to skip blocks of convolutional layers using shortcut connections to form unions called residual blocks. These stacked residual blocks significantly improve training

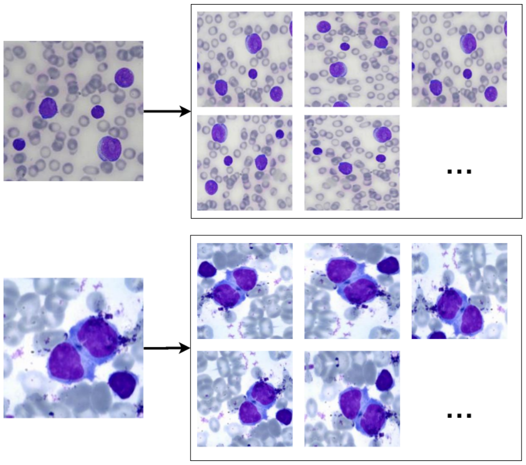


Figure 3 Examples of the results of the data augmentation operations.

efficiency and mainly solve the problem of degradation present in deep networks.

The InceptionV3 [3] architecture emerged as a new version for the GoogLeNet and InceptionV2 architectures. This architecture reduces CNN's complexity in terms of the number of operations performed using Inception modules, which consists 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, it was proposed to perform 1×1 convolutions first, reduce the dimensionality of the characteristics map, and then perform convolutions with the other filters. The use of Inception modules results in a reduction of 28% in the number of parameters compared to traditional convolutional layers.

The Xception [4] 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. It then 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 (filters with size 1 × 1 between channels). Xception achieved superior results when compared to previous versions, despite having fewer layers and parameters. The inclusion of the depthwise separable convolution layers also provided greater efficiency in computational cost, which is less costly and faster than the standard convolution for performing a smaller number of operations.

Based on the AlexNet [33], CaffeNet [34] and VggNet [35] architectures, the authors of Claro et al. cite claro2020convolution developed the Acute Leukemias Recognition Network (Alert Net) which is a CNN for the classification of acute leukemia in blood slides.

Alert Net has five convolutional layers, followed by Batch Normalization and Max Pooling layers. The shallower layers are formed by two fully connected layers, followed by a dropout operation and a softmax layer with three neurons. This model has characteristics existing in sequential architectures presented in state of the art. The authors proposed architectures that would achieve the best compromise between the number of parameters and precision. Therefore, their model with 8 million parameters, is less complex than architectures in the literature.

The authors carried out an ablation study from the initial model to remove or replace layers in Alert Net. Thus, Claro et al. built two models using technologies implemented in some of the CNNs with the best results in the ImageNet competition. They are ResNet [2] and Xception [4]. Developed the Alert Net with a Residual Layer (Alert Net-R) and the Alert Net with Depthwise Separable Convolutions Layer (Alert Net-X).

In Alert Net-R a residual structure similar to ResNet. Initially, max pooling after the input layer is used to resize the original image. This operation's result is concatenated with the maximum result of the second convolutional layer pool. Therefore, the image to be concatenated does not change in the initial convolutional layers. It is observed that the waste generated tends to propagate the essential characteristics of the image during training. Studies presented in the literature prove the efficiency of this approach [2].

The Depth-Wise Separable Convolution layers were introduced in the Xception architecture and provide greater computational efficiency since the number of operations performed during convolution is reduced, so Alert Net-X was developed. That is, they have less complexity and require less training time than regular convolutional layers.

From these architectures developed in the study by Claro et al. [5], we developed a new network architecture called Alert Net-RX. The design for this new architecture involves the insertion of a residual layer in the Alert Net - RX architecture with Depth-Wise Separable Convolution layers. Figure 4 presents the model developed and illustrates the types of layers.

3.4 Ensemble

The ensemble is a technique that existed long before the Deep Learning paradigm emerged [36]. The theory behind this is quite simple and is based on the well-known notion of “wisdom of crowds”: instead of relying on just one model for prediction, a set of multiple (pre-trained) models is created. These models' results are then combined in a final classification by constructing some weighted votes. The original idea was developed to reduce the classifiers' variance to obtain a better overall performance [37].

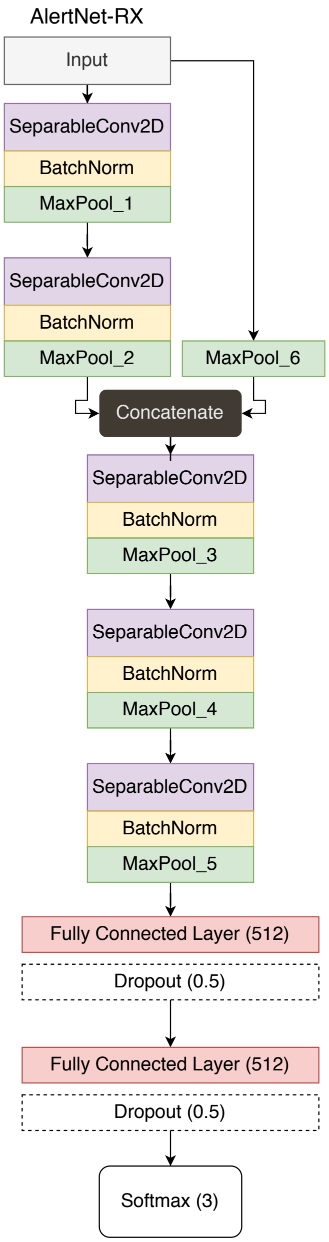


Figure 4 Topologies of the Aler Net - RX a CNN based in Alert Net from [5].

The most commonly used ensemble method for neural networks is weighted average voting. In this vote, later labels are generated by averaging a class's probabilities from the predicted labels, such as the accuracy [38]. Figure 5 presents a general ensemble example.

3.5 Evaluation Metrics

To analyze the classification results, we computed the confusion matrix. Then, from the elements of this matrix, we calculated the accuracy (A), precision (P), recall (R) and F1-score [39].

We also computed the kappa index (k), which is recommended as an appropriate exactitude measure as it can adequately represent the confusion matrix; it takes all elements of the matrix into account, rather than just those on the main diagonal, which occurs when calculating the global classification accuracy [57]. This metric can be calculated as:

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

According to Landis and Koch [40] k assumes values between 0 (zero) and 1 (one). The result is qualified according to the k value as follows: $\mathrm{k} \leq 0.2$: Bad; $0.2 < \mathrm{k} \leq 0.4$: Fair; $0.4 < \mathrm{k} \leq 0.6$: Good; $0.6 < \mathrm{k} \leq 0.8$: Very Good and $\mathrm{k} > 0.8$: Excellent.

The cost function metric (loss) was also used in this work. This function is responsible for saying how far one is from the ideal prediction and, therefore, quantifies the "cost" or "loss" by accepting the prediction generated by the current parameters of the model [41].

4 Experiments

In this section, we present the carried out experiments in this study. The results are structured as follows: Section 4.1 presents the individual results for each CNN mentioned in Section 3.3. Section 4.2 presents the results of the models performing the ensemble experiment. All experiments were applied with the k-fold cross-validation with k equal to 5 and analyzed the results using state of the art metrics.

4.1 Results of Individual CNNs

The CNNs ResNet50, InceptionV3, and Xception presented excellent results in the ImageNet competition. Also, they have a relatively low number of parameters when compared to other sequential architectures. We applied fine-tuning techniques [42], which consists of using a pre-trained architecture to carry out the transfer learning. Thus, training is only conducted in selected layers and with lower learning rates.

The most commonly used fine-tuning technique in the literature is Shallow fine-tuning (SFT) [42]. In SFT, the initial layers are frozen, decreasing the complexity during the CNN training, and only the final layers are retrained. These layers have specific characteristics related to the used dataset. Some authors claim that SFT does not perform well when the target domain differs from one used to pre-train the weights [43]. For example, natural photographic images from ImageNet belong to a different domain compared to blood smear images.

In those situations, it is better to apply Deeply Fine-Tuning (DFT). The DFT approach allows training the entire CNN. However, it requires a higher computational cost and a more considerable amount of data. Table 3 presents the results from obtained applying the SFT and DFT technique with the CNNs found in the literature.

In Table 3, we observed that the performance of the networks using the SFT was lower than using DFT. The ResNet50 architecture, for example, got a kappa value equal to 0 (zero), which means that this CNN classified

Table 3 Results obtained with Shallow Fine Tuning and Deeply Fine Tuning.

Model	**A (%)**	**P (%)**	**R (%)**	**F1-score**	**K**	**Num. param**	**File size**
				Shallow Fine Tuning			
ResNet50	43.56±0.03	18.97±0.02	33.33±0	26.43±0.03	0	24,638,339	98mb
InceptionV3	67.66±2.76	68.85±3.41	67.32±3.29	67.82±2.87	0.5081±0.044	22,853,411	92mb
Xception	68.94±3.85	73.38±2.85	64.42±4.13	67.02±4.38	0.5050±0.060	21,912,107	88mb
				Deeply Fine Tuning			
ResNet50	**96.02**±0.79	**96.07**±0.10	**95.41**±0.75	**96.01**±0.80	**0.9392**±0.012	24,638,339	188mb
InceptionV3	95.44±0.64	95.55±0.52	94.76±0.94	95.42±0.67	0.9298±0.010	22,853,411	175mb
Xception	92.04±1.24	92.338±1.19	90.70±1.50	91.97±1.29	0.8772±0.019	21,912,107	167mb

Table 4 Results obtained by K-fold cross validation for Alert Net variations.

Model	**A (%)**	**P (%)**	**R(%)**	**F1-score(%)**	**K**
Alert Net	93.71±0.68	94.00±0.57	92.62±0.87	93.66±0,70	0.9031±0.010
Alert Net-WD	94.56±0.68	94.69±0.66	93.69±0.89	94.53±0.70	0.9163±0.010
Alert Net-R	93.68±0.84	93.97±0.71	92.67±1.02	93.63±0.88	0.9027±0.013
Alert Net-RWD	**94.74**±0.86	**94.87**±0.79	**93.89**±1.04	**94.72**±0.87	**0.9191**±0.013
Alert Net-X	92.65±1.14	93.04±1.07	91.64±1.28	92.60±1.14	0.8868±0.017
Alert Net-XWD	93.26±1.11	93.39±1.08	92.21±1.31	93.21±1.13	0.8961±0.017
Alert Net-RX	90.89±1.15	91.35±1.04	89.80±1.20	90.83±1.17	0.8597±0.017
Alert Net-RXWD	92.16±0.73	92.57±0.44	91.12±1.00	92.12±0.76	0.8793±0,011

all images in one class. Comparing the results of the CNNs using the SFT and the DFT, one can realize that DFT's use resulted in a substantial performance gain of the pre-trained CNNs. With DFT, ResNet50 obtain with the best results in terms of accuracy (96.02%), precision (94.87%), recall (93.89%), F1-score (94.72%), and kappa (0.9191).

In Table 4, we present the results achieved by architectures proposed by Claro et al. [5] and two new versions called Alert Net-RX and Alert Net-RXWD. In the results, we found that Alert Net-RWD achieved the best results in terms of Accuracy (94.74%), Precision (94.87%), Recall (93.89%), F1-score (94.72%), and kappa (0.9191). We also highlight the results obtained by the Alert Net without dropout.

The results in Table 4 allow us to infer that the dropout layer removal helps achieve better results. The dropout is a regularization technique, and its use reduces the generalization capacity of the model. It would be necessary to increase the model size because typically, the validation dataset's error is much smaller when using dropout, but with accounting, larger models cost more training iterations. When the training dataset is small, the use of the dropout becomes less effective.

4.2 Proposed Ensemble

Figure 5 presents our ensemble approach. It consists of four CNNs: three pre-trained Resnet50, InceptionV3, Xception, and Aler Net-RWD. The results presented in this section demonstrate the reason for choosing these four networks.

For the development of the proposed CNNs Ensemble, we evaluated the literature's architecture and the models generated by Claro et al. [5]. The predictions of the pre-trained CNN models are calculated to build an ensemble model. An advantage of using this model set is that it does not need training, since the average of the predictions does not require any parameter to be learned.

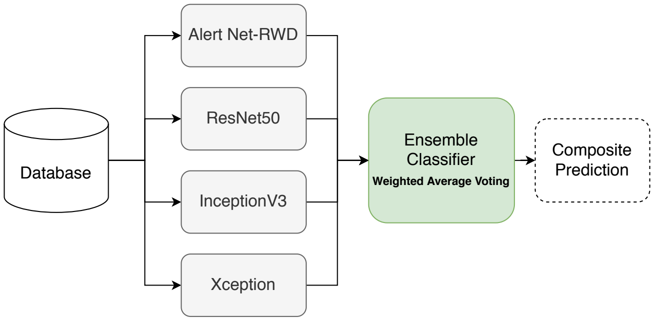


Figure 5 Proposed ensemble flowchart.

First of all, we evaluated the Alert Net variations, since they were proposed to identify acute leukemias. Table 5 presents the results achieved through these ensembles. The former ensemble was carried out with the four CNNs with a dropout layer (Alert Net, Alert Net-R, Alert Net-X, and Alert Net-RX). The second ensemble was formed by the CNNs withouth a dropout layer. Finally, the third ensemble is composed of the eight Alert Net architectures. We can infer that the set using only as networks without a dropout layer obtained a better performance.

Table 6 illustrates the results of the prediction of ensemble models with pre-trained CNNs. In the first row, the ensemble was formed by the refined CNNs Resnet50, InceptionV3, and Xception. We chose these CNNs since they achieved the best-ranking performance using Deeply Fine Tuning.

To improve the ensemble, we carried out tests, including the eight versions of Alert Net separately. In the end, we found that the best set of CNNs to compose the ensemble is the one shown in the second line of Table 6 . That is the union of the pre-trained networks Resnet, Inception, and Xception with Alert Net-RWD.

Table 5 Results achieved with different ensembles formed by Alert Net variations.

Model Ensemble	A (%)	P (%)	R(%)	F1-score(%)	K
Alert Net, Alert Net-R, Alert Net-X, Alert-RX	93.31±0.95	93.90±0.84	93.30±0.94	93.33±0.96	0.8971±0.014
Alert Net-WD, Alert Net-RWD, Alert Net-XWD, Alert Net-RXWD	**94.50**±0.83	**94.82**±0.63	**94.52**±0.84	**94.51**±0.82	**0.9153**±0.013
Alert Net, Alert Net-WD, Alert Net-R, Alert Net-RWD, Alert Net-X, Alert Net-XWD, Alert Net-RX, Alert Net-RXWD	94.22±0.73	94.83±0.46	94.23±0.71	94.26±0.71	0.9111±0.011

Table 6 Results achieved with different ensembles formed by pre-trained CNNs.

Model Ensemble	A (%)	P (%)	R (%)	F1-score (%)	K
Resnet50, InceptionV3, Xception	95.38±0.88	95.71±0.59	95.39±0.87	95.40±0,83	0.9288±0,013
Alert Net-RWD, Resnet50, InceptionV3, Xception	**96.17±0,56**	**96.38±0,43**	**95.94±0,55**	**96.18±0,54**	**0.9411±0,008**

CNNs suffer from the limitation of high variance, as they are highly dependent on the specifications of the training data and prone to overfitting, which reduces their generalizability. We address this problem by training various models to obtain a diverse set of predictions that, when combined, can provide a set of viable solutions.

It is observed that for the construction of an ensemble, it is important to select diversified CNNs that present high precision rates in several regions in the characteristics space. For this reason, we evaluate model combinations and determine the best one to build the proposed ensemble. The 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.

5 Discussion

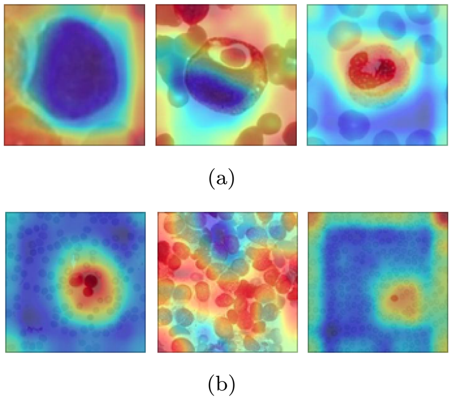
We compare the proposed ensemble results and the ones obtained by literature works that address the same problem. From Table 7, one can realize that state of the art presented higher accuracy values than the proposed approach. However, the number of images used in those studies is at least four times less than the number of images used in this work. Another essential point to be highlighted is related to the images' heterogeneity, as they come from eighteen different datasets. This characteristic leads to a greater diversity in the training data, which leads to the achievement of a robust method for different input image types.

Figure 6 shows examples of the Alert Net-RWD activation maps for the three classes. It is possible to identify which regions are used to differentiate healthy images from those with acute leukemia (lymphoid or myeloid).

The number of leukocytes may vary depending on the input image. We see in Figure 6 that Alert Net-RWD generates different activation map patterns for each of these situations. Also, as it is trained in different databases, the proposed model can adapt to different characteristics, it is possible to observe that the maps differ from one base to another. However, CNN activates different regions for each class. For example, in Figure 5.a, of the ALL class, the leukocyte turned predominantly blue. This pattern changes in the other classes, becoming mostly red in the HBS class.

Table 7 Comparison among the results obtained by the proposed ensemble against the ones obtained by related methods.

Method	Descritors	N. of images	Accuracy
Rawat et al. [18]	Geometrical, color and texture	420	99.50%
Laosai and Chamnongthai [19]	Shape, color, Texture and number of nucleoli	500	99.85%
Tran et al. [20]	Deep Features	1,636	97.30%
Claro et al. [5]	Deep Features	2,415	97.18%
Proposed method	Deep Features	3,293	96.17%


(a)

(b)

Figure 6 Examples of activation maps for blood slides, (a) images with one leukocyte, (b) images with various leukocytes. The first column are images of the ALL class; the second column are images of the AML class and the third of the HBS class.

6 Conclusion

The conducted systematic survey showed that many researchers had focused their efforts on the Computer-Aided Diagnostic systems field, where the automatic diagnosis of leukemia can be found.

This study proposed an ensemble model with CNNs for the automated diagnosis of acute lymphoid leukemia and acute myeloid leukemia. The results achieved in 3,293 images are encouraging. The literature works showed results superior to ours; however the quantity and diversity of images applied in our work are superior to that of the compared works, thus increasing the method's robustness.

Another essential point to be highlighted is the use of the ensemble method, where it reduces the model's standard deviation rate, ideally combining the predictions of various models. The ensemble model's performance simulates real-world conditions with standard deviation and reduced overfitting, leading to improved generalization. We believe that the proposed results are beneficial for developing clinically valuable solutions to detect and differentiate images of acute leukemia in blood slides.

For future approaches, the proposed model needs to be applied to a more significant number of images. Moreover, in addition to differentiation of the three classes proposed in this work, a distinction will also be made between the images that have Chronic Lymphocytic Leukemia and Chronic Myeloid Leukemia.

References

[1] Gregory S. Travlos. Normal structure, function, and histology of the bone marrow. *Toxicologic Pathology*, 34(5):548–565, 2006.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1800–1807, 2017.

[5] Maíla Claro, Luis Vogado, Rodrigo Veras, André Santana, João Tavares, Justino Santos, and Vinicius Machado. Convolution neural network models for acute leukemia diagnosis. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 63–68. IEEE, 2020.

[6] Luis H. S. Vogado, Rodrigo M. S. Veras, Flávio H. D. Araújo, Romuere Rôdrigues Veloso e Silva, and Kelson Rômulo Teixeira Aires. Leukemia diagnosis in blood slides using transfer learning in cnns and SVM for classification. *Engineering Applications of Artificial Intelligence*, 72:415–422, 2018.

[7] Mohamed Loey, Mukdad Naman, and Hala Zayed. Deep transfer learning in diagnosing leukemia in blood cells. *Computers*, 9(2):29, 2020.

[8] Ivan Vincent, Ki-Ryong Kwon, Suk-Hwan Lee, and Kwang-Seok Moon. Acute lymphoid leukemia classification using two-step neural network classifier. *Frontiers of Computer Vision*, pages 1–4, 2015.

[9] Vanika Singhal and Preety Singh. *Texture Features for the Detection of Acute Lymphoblastic Leukemia*, pages 535–543. Springer Singapore, Singapore, 2016.

[10] T. T. P. Thanh, Caleb Vununu, Sukhrob Atoev, Suk-Hwan Lee, and Ki-Ryong Kwon. Leukemia blood cell image classification using convolutional neural network. *International Journal of Computer Theory and Engineering*, 10(2):54–58, 2018.

[11] Sarmad Shafique and Samabia Tehsin. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technology in Cancer Research and Treatment*, 17:1–7, september 2018.

[12] Amjad Rehman, Naveed Abbas, Tanzila Saba, Syed Ijaz ur Rahman, Zahid Mehmood, and Hoshang Kolivand. Classification of acute lymphoblastic leukemia using deep learning. *Microscopy Research and Technique*, pages 1–8, October 2018.

[13] Tatdow Pansombut, Siripen Wikaisuksakul, Kittiya Khongkraphan, and Aniruth Phon-on. Convolutional neural networks for recognition of lymphoblast cell images. *Computational Intelligence and Neuroscience*, 2019:1–12, 2019.

[14] Nizar Ahmed, Altug Yigit, Zerrin Isik, and Adil Alpkocak. Identification of leukemia subtypes from microscopic images using convolutional neural network. *Diagnostics*, 9(3):104, 2019.

[15] Shiv Gehlot, Anubha Gupta, and Ritu Gupta. Sdct-auxnetθ: Dct augmented stain deconvolutional cnn with auxiliary classifier for cancer diagnosis. *Medical Image Analysis*, 61:101661, 2020.

[16] Monica Madhukar, Sos Agaian, and Anthony T Chronopoulos. Deterministic model for acute myelogenous leukemia classification. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 433–438. IEEE, 2012.

[17] D Goutam and S Sailaja. Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 1–5. IEEE, 2015.

[18] Jyoti Rawat, Annapurna Singh, Bhadauria HS, Jitendra Virmani, and Jagtar Singh Devgun. Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia. *Biocybernetics and Biomedical Engineering*, 37(4):637 – 654, 2017.

[19] Jakkrich Laosai and Kosin Chamnongthai. Classification of acute leukemia using medical-knowledge-based morphology and cd marker. *Biomedical Signal Processing and Control*, 44:127–137, 2018.

[20] Thanh Tran, Jin-Hyuk Park, Oh-Heum Kwon, Kwang-Seok Moon, Suk-Hwan Lee, and Ki-Ryong Kwon. Classification of leukemia disease in peripheral blood cell images using convolutional neural network. *Journal of Korea Multimedia Society*, 21(10):1150–1161, 2018.

[21] Ruggero Donida Labati, Vincenzo Piuri, and Fabio Scotti. All-idb: The acute lymphoblastic leukemia image database for image processing. In *IEEE International Conference on Image Processing*, pages 2045–2048, 2011.

[22] Omid Sarrafzadeh and Alireza Mehri Dehnavi. Nucleus and cytoplasm segmentation in microscopic images using k means clustering and region growing. *Advanced Biomedical Research*, pages 79–87, December 2015.

[23] Marian. Rollins-Raval, Jay. Raval, and Lydia. Contis. Experience with cellavision dm96 for peripheral blood differentials in a large multi-center academic hospital system. *Journal of Pathology Informatics*, 3(29):1–9, 2012.

[24] Omid Sarrafzadeh, Hossein Rabbani, Ardeshir Talebi, and Hossein Usefi Banaem. Selection of the best features for leukocytes classification in blood smear microscopic images. In *SPIE Medical Imaging*, volume 9041, 2014.

[25] Omid Sarrafzadeh, Hossein Rabbani, Alireza Mehri Dehnavi, and Ardeshir Talebi. Detecting different sub-types of acute myelogenous leukemia using dictionary learning and sparse representation. In *IEEE International Conference on Image Processing*, pages 3339–3343. IEEE, 2015.

[26] Monica Madhukar, Sos Agaian, and Anthony T. Chronopoulos. Automated screening system for acute myelogenous leukemia detection in blood microscopic images. *IEEE Systems Journal*, 8(3):995–1004, 2014.

[27] Alessandra Mendes Pacheco Guerra Vale, Ana Maria Guimarães Guerreiro, Adrião Duarte Dória Neto, Geraldo Barroso Cavalvanti Junior, Victor Cezar Lucena Tavares de Sá Leitão, and Allan Medeiros Martins. Automatic segmentation and classification of blood components in microscopic images using a fuzzy approach. *Revista Brasileira de Engenharia Biomédica*, 30:341–354, 2014.

[28] J. Böhm. Pathologie-websites im world wide web. *Der Pathologe*, 29(3):231–242, 2008.

[29] Xin Zheng, Yong Wang, Guoyou Wang, and Zhong Chen. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.

[30] Rahul Duggal, Anubha Gupta, Ritu Gupta, and Pramit Mallick. Sd-layer: Stain deconvolutional layer for cnns in medical microscopic imaging. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention*, pages 435–443. Springer, 2017.

[31] Seyed Hamid Rezatofighi and Hamid Soltanian-Zadeh. Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics*, 35(4):333 – 343, 2011.

[32] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, July 2019.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[34] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[35] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[36] Belur V Dasarathy and Belur V Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, 1979.

[37] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[38] Ali Yazdizadeh, Zachary Patterson, and Bilal Farooq. Ensemble convolutional neural networks for mode inference in smartphone travel survey. *IEEE Transactions on Intelligent Transportation Systems*, 21(6):2232–2239, 2019.

[39] David MW Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation, school of informatics and engineering, flinders university, adelaide, australia. Technical report, TR SIE-07-001, Journal of Machine Learning Technologies 2: 1 37-63., 2007.

[40] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[41] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv:1702.05659*, 2017.

[42] Kitsuchart Pasupa and Wisuwat Sunhem. A comparison between shallow and deep architecture classifiers on small dataset. In *International Conference on Information Technology and Electrical Engineering*, pages 1–6. IEEE, 2016.

[43] Mohammadhassan Izadyyazdanabadi, Evgenii Belykh, Michael Mooney, Nikolay Martirosyan, Jennifer Eschbacher, Peter Nakaji, Mark C. Preul, and Yezhou Yang. Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization for neurosurgical cle images. *Journal of Visual Communication and Image Representation*, 54:10–20, 7 2018.