Computer Methods and Programs in Biomedicine 202 (2021) 105999

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

# A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images

Laura Boldú[a], Anna Merino[a,1,*], Andrea Acevedo[a,b], Angel Molina[a], José Rodellar[b]

[a] *Hospital Clínic de Barcelona-IDIBAPS, Haematology and Cytology Unit, CORE Laboratory, Biomedical Diagnostic Centre, Spain*
[b] *Technical University of Catalonia, Barcelona East Engineering School, Department of Mathematics, Spain*

ARTICLE INFO

*Article history:*
Received 31 July 2020
Accepted 8 February 2021

*Keywords:*
Leukemia
Deep learning
Convolutional neural networks
Morphological analysis
Blood cell automatic recognition

ABSTRACT

*Background and objectives:* Morphological differentiation among blasts circulating in blood in acute leukaemia is challenging. Artificial intelligence decision support systems hold substantial promise as part of clinical practise in detecting haematological malignancy. This study aims to develop a deep learning-based system to predict the diagnosis of acute leukaemia using blood cell images.

*Methods:* A set of 731 blood smears containing 16,450 single-cell images was analysed from 100 healthy controls, 191 patients with viral infections and 148 with acute leukaemia. *Training and testing sets* were arranged with 85% and 15% of these smears, respectively. To find the best architecture for acute leukaemia classification VGG16, ResNet101, DenseNet121 and SENet154 were evaluated. Fine-tuning was implemented to these pre-trained CNNs to adapt their layers to our data. Once the best architecture was chosen, a system with two modules working sequentially was configured (ALNet). The first module recognised abnormal promyelocytes among other mononuclear blood cell images, such as lymphocytes, monocytes, reactive lymphocytes and blasts. The second distinguished if blasts were myeloid or lymphoid lineage. The final strategy was to predict patients' initial diagnosis of acute leukaemia lineage using the blood smear review. ALNet was assessed with smears of the *testing set*.

*Results:* ALNet provided the correct diagnostic prediction of all patients with promyelocytic and myeloid leukaemia. Sensitivity, specificity and precision values of 100%, 92.3% and 93.7%, respectively, were obtained for myeloid leukaemia. Regarding lymphoid leukaemia, a sensitivity of 89% and specificity and precision values of 100% were obtained.

*Conclusions:* ALNet is a predictive model designed with two serially connected convolutional networks. It is proposed to assist clinical pathologists in the diagnosis of acute leukaemia during the blood smear review. It has been proved to distinguish neoplastic (leukaemia) and non-neoplastic (infections) diseases, as well as recognise the leukaemia lineage.

## 1. Introduction

Acute leukaemia are a heterogeneous group of blood-related cancers characterized by the abnormal proliferation of blast cells in bone marrow, where causes a replacement of normal cells and a decrease of the three haematopoietic lines in peripheral blood. They represent the 11th and 10th most frequent cause of cancer occurrence and death worldwide, respectively, with more than 300,000 deaths estimated in 2018. Acute myeloid leukaemia is an aggressive cancer with 3.7 new cases per 100,000 habitants per year and with a 5-year survival of only 19% in Europe [1]. A timely and accurate diagnosis is crucial for an effective disease management.

The World Health Organization (WHO) considers morphology, along with other complementary tests such as immunophenotype, cytogenetic and molecular biology, essential for the integrated diagnosis of acute leukaemia [2]. This is why the starting point in their diagnosis is still the detection of blasts in blood. Nevertheless, smear review is time consuming, requires well-trained personnel and is prone to intra-observer variability, which is particularly true when dealing with blasts. Indeed, subtle interclass morphological differences exist for leukaemia types, which turns into low specificity scores in the routine screening [3]. They are well-known the difficulties that clinical pathologists have in the discrimination among different blasts and the subjectivity associated with

[*] Corresponding author.
*E-mail address:* amerino@clinic.cat (A. Merino).
[1] Postal address: Villarroel Street, 170. ZIP Code: 08036 City: Barcelona. Country: Spain.

https://doi.org/10.1016/j.cmpb.2021.105999
0169-2607/© 2021 Elsevier B.V. All rights reserved.

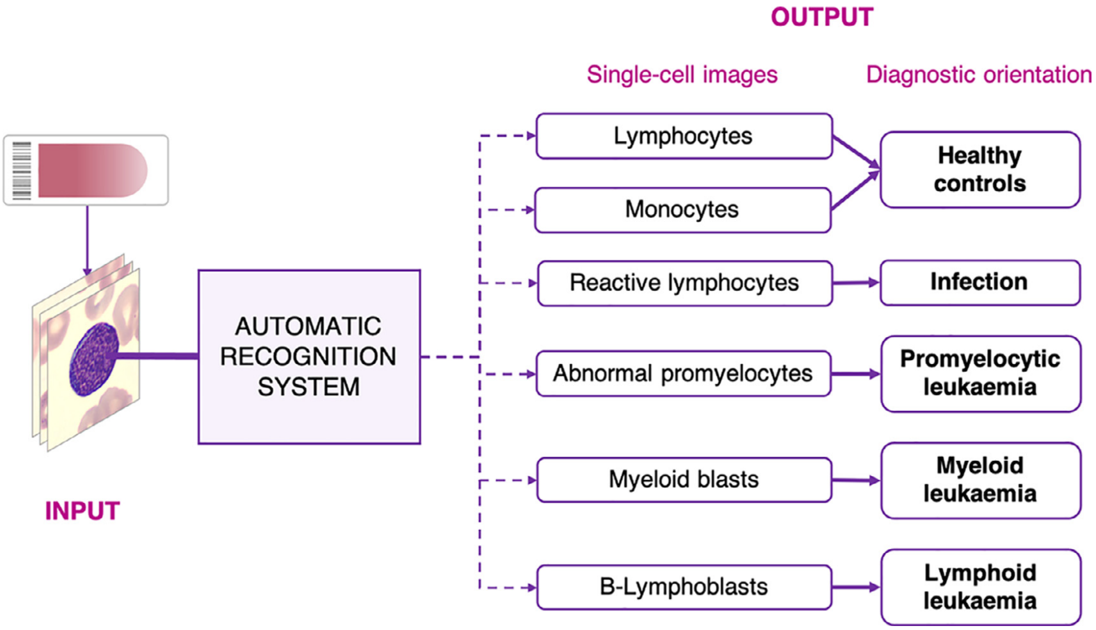

**Fig. 1.** Scheme of the automatic classification system proposed as a support diagnostic tool for the diagnostic prediction of: 1) infection, 2) acute promyelocytic leukaemia, 3) acute myeloid leukaemia and 4) acute lymphoid leukaemia.

their morphological recognition. Identifying the leukaemia lineage is crucial since the prognostic and immediate therapeutic consequences drastically depend on this differentiation. This challenging problem has been scarcely addressed in the literature in spite that it is well known that automated blood cell image analysers tend to underestimate the number of blast cells [4,5].

Image analysis, quantitative morphological features and machine learning approaches have been the main technological tools adopted in the last decade to overcome such drawbacks [6,7]. The late explosion of deep learning has shifted the focus towards new classification models using convolutional neural networks (CNNs) [7,8]. The application of these automatic classification systems will increasingly become a part of clinical practise in the coming years in the field of haematological malignancy [9,10].

Recently, CNNs have proven to be successful for the classification of normal white blood cells [11–16], the distinction of specific cell subtypes, such as erythroid and myeloid precursors [17], and also for the diagnosis of lymphoma from lymph node digital images [18]. In the context of the morphological classification of cells circulating in peripheral blood in acute leukaemia, most of the previous works used CNNs with transfer learning techniques [19–26], while others used them for feature extraction and afterwards classify images with other well-known traditional machine learning techniques such as support vector machine [27,28].

Literature reveals that lymphoid leukaemia has received more attention, as the automatic recognition of acute leukaemia with CNNs has mainly been addressed in two cases: 1) to differentiate lymphoblasts and leukocytes with very diverse cell morphology, including neutrophils, eosinophils, basophils and lymphocytes [19,24,26,27]; and 2) to separate lymphoblast subtypes [20,21]. Furthermore, CNNs have also been employed in bone marrow images in two problems: 1) the differentiation of the subtypes of acute lymphoid leukaemia regarding the French-American-British (FAB) classification [29], which has recently been replaced by the WHO classification for clinical practise; and 2) the discrimination between acute and chronic leukaemia [30]. It is remarkable that none of these previous works have considered the presence of reactive lymphocytes, which could be difficult since these lymphoid cells share some morphological similarities with blasts [31,32], nor the distinction among blasts from different origin [2,33]. Moreover, the detection of abnormal promyelocytes is of utmost importance because acute promyelocytic leukaemia patients can suffer severe bleeding and die if no treatment is initiated promptly [34].

The objective of this paper is to develop a new CNN-based system (ALNet) to predict the initial diagnostic orientation of acute leukaemia using blood cell images, which must be highly sensitive and specific for clinical application. The system input is a set of cell images of an individual patient's smear, and the output is the prediction of one of the following diagnoses: acute promyelocytic leukaemia (APL), acute myeloid leukaemia (AML non-APL), acute lymphoid leukaemia (ALL) or infection.

The recognition of lymphoblasts from myeloblasts by morphology in the blood smear review is a challenging problem because they share morphological characteristics. To the authors' knowledge, it is the first time in the literature that reactive lymphocytes, abnormal promyelocytes alongside myeloid blasts and B-lymphoblasts have been jointly considered for their automatic recognition using CNN approaches. A previous publication by the authors [35] addressed the recognition of these cell groups using a linear discriminant analysis classifier in which the overall accuracy was 85.8%. Our new contribution is clinically relevant since the proposed sequential system ALNet is able to distinguish neoplastic (leukaemia) and non-neoplastic (infections) diseases, as well as recognise the leukaemia lineage with an overall accuracy of 94.2%. This improvement makes possible its integration as a decision support system to assist pathologists when there is suspicion of acute leukaemia.

## 2. Material and methods

### 2.1. Overview

The main objective of this study is to design an automatic classification system to work within the general scheme shown in Fig. 1 and based on CNNs. The rationale for this scheme is twofold: 1) the first step in the diagnosis of acute leukaemia is the morphological review of the peripheral blood smear; and 2) the clinical pathologist could select those abnormal cell images responsible for

**Table 1**
Diagnosis of the patients in the respective sets of study: acute leukaemia (AL) according to the WHO 2016 classification and viral infections. The table also indicates the diagnosis and AL subtype with results of the most relevant complementary tests. ALL-B acute lymphoid leukaemia; AML, acute myeloid leukaemia; APL, acute promyelocytic leukaemia; HLA-DR; human leucocyte antigen; MPO, myeloperoxidase; NOS, not otherwise specified; P, number of patients.

| DIAGNOSIS | AL SUBTYPE AND COMPLEMENTARY TESTS | P | Training N° images | Testing N° images |
|---|---|---|---|---|
| AML with recurrent genetic abnormalities | APL with *PML-RARA* (HLA-DR-, CD34-, CD13+, CD33+) | 15 | 2,575 | 1,358 |
| | AML with t(6;9) (p23;q34.1); *DEK-NUP214* (CD45 weak, CD34+, HLA-DR weak, CD117 weak, CD13+, CD33+) | 5 | 144 | 65 |
| | AML with inv(16) (p13.1q22) or t(16;16) (p13.1;q22); *CBFB-MYH11* (CD45 weak, CD117+, MPO+, CD13+, CD33+, CD4+) | 1 | 21 | 12 |
| | AML with t(8;21) (q22;q22.1); *RUNX1-RUNX1T1* (CD45 weak, HLA-DR+, CD34+, CD117+, MPO weak, CD13+, CD33+) | 2 | 47 | 29 |
| | AML with inv(3) (q21.3q26.2) (CD45 weak, CD34+, HLA-DR+, CD117+, CD13 weak, MPO weak) | 2 | 62 | 25 |
| | AML with t(9;11) (p21.3;q23.3); *MLLT3-KMT2A* | 5 | 138 | 41 |
| | AML with mutated *NPM1* (HLA-DR+, CD117+, CD13+, CD33+, CD123+) | 13 | 669 | 68 |
| AML with myelodysplasia-related changes | CD45 weak, CD34+, HLA-DR+, CD117+, CD13+, CD33+, CD123+ | 51 | 1,153 | 390 |
| AML, NOS | Acute monoblastic/monocytic leukaemia (HLA-DR+, CD13+, CD33+, CD64+, CD4+, CD36+, CD11b+) | 17 | 636 | 144 |
| ALL-B/lymphoma with recurrent genetic abnormalities | ALL-B/lymphoma with t(9;22) (q34.1;q11.2); *BCR-ABL1* (CD123+, CD19+, CD22+, CD79a+) | 9 | 141 | 337 |
| ALL-B/lymphoma, NOS | CD19+, CD22+, CD79a+, CD10+, CD20+, CD34+ | 13 | 199 | 285 |
| | CD45 weak, CD19+, CD79a+, TdT+, CD10-, CD20- | 15 | 277 | 401 |
| Viral or other infections | | 191 | 1,852 | 459 |
| TOTAL | | 339 | 7,914 | 3,614 |

**Table 2**
Total number of blood cell images and smears used in this work. Images and smears are grouped by class for each dataset and the number of healthy controls and patients for each entity is provided.

| | | Number of images | | Number of smears | | Number of controls and patients |
|---|---|---|---|---|---|---|
| Entity | Cell type | Training | Testing | Training | Testing | |
| **Healthy controls** | **Lymphocytes** | 3,288 | 312 | 82 | 13 | 100 |
| | **Monocytes** | 1,205* | 117 | 83 | 4 | |
| **Infections** | **Reactive lymphocytes** | 1,852* | 459 | 168 | 46 | 191 |
| **Acute leukaemia** | **Abnormal promyelocytes** | 2,575 | 1,358 | 19 | 14 | 148 |
| | **Myeloid** **Myeloblasts** | 2,234 | 630 | 200 | 25 | |
| | **blasts** **Monoblasts** | 636 | 144 | 24 | 5 | |
| | **B-Lymphoblasts** | 617* | 1,023 | 48 | 27 | |
| | **TOTAL** | 12,407 | 4,043 | 623 | 108 | 439 |

* Cell groups with less than 2,500 images for training, which were up-sampled with data augmenting techniques to balance all the classes

suspecting of acute leukaemia and upload them into the system to obtain the assistance of an automatic classification.

Our design is focused on achieving the following specifications:

1 The prompt and accurate detection of promyelocytic leukaemia, motivated by the emergency of avoiding severe bleeding and disseminated intravascular coagulation with patients' death.
2 The discrimination between myeloid and lymphoid leukaemia lineages, motivated by the different strategical therapies they require.

The system development and assessment involve two main stages:

1 In the first stage, we develop a CNN model to automatically classify images of six cell groups: 1) lymphocytes, 2) monocytes, 3) reactive lymphocytes, 4) abnormal promyelocytes, 5) myeloid blasts (myeloblasts and monoblasts) and 6) B-lymphoblasts.
2 Based on the results obtained when testing this classifier, in the second stage we design and evaluate a recognition system where the input will be a set of cell images of an individual patient's blood smear, and the output will be the prediction of one of the following diagnoses: APL, AML, ALL or infection.

The remaining of this section will be devoted to describe the most relevant issues of the CNN classifier development, including the used image database. To keep a continuous thread in the presentation, all the details concerning the second stage will be presented in Section 3.

### 2.2. Image datasets

Blood samples, collected in EDTA, were automatically prepared using the slide maker-stainer SP10i (Sysmex, Kobe, Japan) and stained with May Grünwald-Giemsa. Digital images of blood cells were acquired by the CellaVision®DM96 (CellaVision, Lund, Sweden) (363 × 360 pixels) from smears compiled during daily work in the Core Laboratory of the Hospital Clinic of Barcelona. Each digital image contained a single cell. Cell images were identified by pathologists according to their morphological characteristics. Patients' diagnoses were confirmed by the integration of all supplementary data: clinical data, morphology, flow cytometry, cytogenetics and molecular biology. Patients with acute leukaemia were diagnosed by the Clinic Haematology Service of the Hospital following the WHO 2016 classification [36]. We regard these confirmed diagnoses as the ground truth for training and evaluating the models. Table 1 shows the number of images and patients with acute leukaemia and infections included in *training* and *testing sets*.

Table 2 details the number of images and smears corresponding to the different cell classes included in *training* and *testing sets* and the number of healthy controls and patients from which they were obtained. Images and smears are grouped in six classes for each dataset: lymphocytes, monocytes, reactive lymphocytes, abnormal promyelocytes, myeloid blasts [myeloblasts and monoblasts) and

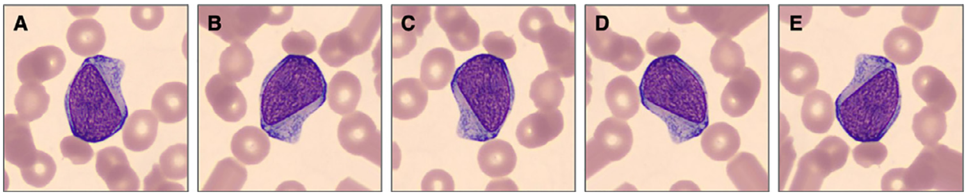

**Fig. 2.** Examples of some of the transformations applied to increase the number of images of the *train set*. (A) Original image of a B-lymphoblast; (B, C, D, E) Versions of the same image rotated and flipped.

B-lymphoblasts. Smears were obtained from 100 healthy controls, 191 patients with viral infections and 148 patients with acute leukaemia.

The *training set* was arranged with 85% of the smears [623 with 12,407 images), distributed into the above groups of interest. The *testing set,* reserved for the final assessment of the classification system, included the remaining 15% of the smears [108 with 4,043 images).

Data augmenting was performed to balance the *training set* by incrementing the number of images and obtaining the same number per each cell class. Individual images were modified applying randomly transformations as vertical and horizontal flips and rotations from 0 to 60 degrees [37]. Up to four different versions of each original image were obtained from those groups containing less than 2,500 images (monocytes, reactive lymphocytes and B-lymphoblasts) as illustrated in Fig. 2.

### 2.3. Development and testing of the first CNN classifier

CNNs are multi-layered architectures able to automatically extract complex and high-dimensional features from a large set of images [38]. Through scanning images, they detect and learn patterns by the interaction of their elementary units (neurons) analogising the learning process of human brain. Unlike traditional computer vision models where hand-crafted features are extracted and used to train a classifier, CNNs can learn how to extract relevant features and use them for classification purposes. Moreover, in contrast to previous machine learning methods, the automatic classification of blood cells will not explicitly rely on complex segmentation of the cell regions of interest and the further feature selection.

The first element of a network is always the input layer, which reads image's pixels. Following it are convolutional layers, which are composed by a set of filters responsible for detecting specific patterns in images and extract features. Its output is a set of feature maps. Depending on its architecture, CNNs could contain one or consecutive convolutional layers creating convolutional blocks. After each convolutional layer or block, it usually comes a pooling layer, which reduces the dimension of feature maps while preserving the relevant information and eliminating irrelevant details. Max pooling function is the most widely used, which reduces feature maps size by taking the maximum values. This combination of convolutional and pooling layers is being repeated throughout all the network structure to extract more complex features each time. The depth of a network depends on how many repetitions of these layers are. At the end of them, a set of feature maps representative of the input images is obtained.

In this work, we adopted the concept of fine-tuning, which consists on modifying a CNN architecture previously trained for another task and re-train and adapt some of its layers with our new data to obtain an end-to-end classifier [39]. Its purpose was to take advantage of the knowledge of a CNN pre-trained with a large dataset, which helped to overcome the deficit of training examples and served as an effective weight initialization to, afterwards, use it for the classification of the images of our study. This approach is usually implemented with CNNs with large quantity of parameters because training them from scratch can affect their ability to generalize and may result into low accuracies or overfitting [40].

To examine the effects of different CNN frameworks on model performance, we investigated four well-known CNNs already pre-trained with the ImageNet database [41]: VGG16, ResNet101, DenseNet121 and SENet154 [42–45]. VGG16, ResNet101 and SENet154 were chosen based on their high performance obtained in previous studies by our group. We achieved an overall assessment accuracy of 96.2% with a fine-tuned VGG16 for the recognition of eight groups of normal blood cells [16]. We used a pre-trained ResNet101 to differentiate up to 11 cell groups including normal, reactive, abnormal lymphocytes and blasts, achieving an overall accuracy of 82.8% [46]. We also implemented a SENet154 for the diagnostic prediction among lymphoma, acute leukaemia and infection with an overall assessment accuracy of 94.5% [47]. DenseNet121 was chosen from literature because of its 95.3% accuracy for predicting leukaemia diagnosis [30].

For the transfer learning approach, we removed their last layers, which were fully connected layers trained for classifying 1,000 categories. Afterwards, two fully connected layers were coupled at the end of each network. Fully connected layers learn how to combine the feature maps to perform the final classification of the input images. The design of these layers for each evaluated architecture is detailed in the following paragraph.

A total of 512 feature maps were obtained from VGG16 and converted to an array of 25,088 features, which were the input to the two fully connected layers, the first layer with 1,024 nodes and the second with 512 nodes. From ResNet101 and SENet154, a number of 2,048 feature maps were obtained and converted to an array of 100,352 features. These features were fed to a first fully connected layer of 4,096 nodes and the output to the second of 512 nodes. Regarding DenseNet121, a total of 1,024 feature maps were obtained and converted to an array of 50,176 features. Its first fully connected layer had 2,048 nodes and the second had 512 nodes. To obtain the final classification, a third fully connected layer with six nodes was configured for all the four architectures, one node for each cell class (see Fig. 1). This last layer predicted the class of the input images as probability values by applying a softmax operation.

Training a neural network is an iterative process that typically consists of two phases: a forward and a backpropagation. In the first one, all input images are passed through the network at each iteration (epoch). Once each image is classified, a loss function is calculated by the difference (error) between the prediction made by the network and the label assigned by the clinical pathologist according to the ground truth, which is the diagnosis based on all complementary information following the WHO 2016 (see Table 1) [36]. Loss increases when the predicted probability diverges from the true label. Afterwards, in the second phase, the error is propagated backwards through the network and all the weights are updated to minimize the loss function and obtain a high accuracy (rate of images correctly classified). This two-phase process usually needs to be repeated for several epochs to obtain an optimum trained model.

**Table 3**
Overall accuracies of the pre-trained CNNs when changing the number of convolutional blocks to be trained with 470 iterations and with the two evaluation approaches: hold-out and 5-fold cross-validation. For the hold-out approach, validation and testing accuracies are shown. For the 5-fold cross validation, we present the mean and the standard deviation of the testing accuracy computed among the five folds. The last two columns show the training time (in minutes) achieved when fine-tuning the entire convolutional blocks, and the classification time (in seconds) when evaluating the selected models with the respective *testing sets*.

| | | Number of convolutional blocks fine-tuned | | | | | Testing | Training time (min) | Class. time (s) |
|---|---|---|---|---|---|---|---|---|---|
| | Model | 1 | 2 | 3 | 4 | Whole | | | |
| **Hold-out** | **VGG16** | 89.2% | 88.0% | 88.2% | 88.6% | 94.6% | **88.4%** | 6.78 | 12 |
| | **ResNet101** | 90.1% | 89.7% | 90.3% | 90.2% | 93.3% | 84.0% | 12.52 | 22.4 |
| | **DenseNet121** | 84.4% | 86.8% | 87.5% | 88.0% | 93.6% | 84.8% | 7.9 | 14 |
| | **SENet154** | 89.3% | 89.8% | 89.8% | 88.9% | 94.6% | **85.0%** | 48.2 | 41 |
| **5-fold cross validation** | **VGG16** | 76.3%±0.93 | 76.5%±0.67 | 76.2% ±0.97 | 76.1% ±1.36 | **86.9%**±0.68 | | 123.1 | 29.5 |
| | **ResNet101** | 80.8%±1.99 | 81.2%±0.54 | 80.2% ±0.56 | 81.8% ±0.41 | 86.8%±0.62 | | 136.2 | 32 |
| | **DenseNet121** | 82.6%±1.10 | 82.9%±1.33 | 82.4% ±1.20 | 83.1% ±1.19 | 86.3%±0.93 | | 262.5 | 63 |
| | **SENet154** | 79.5%±0.72 | 80.2%±0.83 | 80.5% ±0.58 | 80.3% ±0.60 | **87.2%**±1.22 | | 426.7 | 81.5 |

**Table 4**
Confusion matrix of the classification results (in %) of the VGG16 for the images of the *testing set*. Rows indicate the true class and columns represent the predicted class supplied by the network. Diagonal values are the true positive rates for each cell type. The overall classification accuracy was 88.4%.

| | | Predicted class | | | | | |
|---|---|---|---|---|---|---|---|
| | | Lymphocytes | Monocytes | Reactive lymphocytes | Abnormal promyelocytes | Myeloid blasts | B-Lymphoblasts |
| True class | **Lymphocytes** | **98** | 0 | 1 | 0 | 0 | 1 |
| | **Monocytes** | 0 | **91** | 3 | 1 | 4 | 1 |
| | **Reactive lymphocytes** | 0 | 0 | **97** | 0 | 2 | 1 |
| | **Abnormal promyelocytes** | 0 | 3 | 0 | **94** | 3 | 0 |
| | **Myeloid blasts** | 0 | 2 | 0 | 2 | **87** | 9 |
| | **B-Lymphoblasts** | 0 | 0 | 0 | 3 | 22 | **75** |

In our work, cross-entropy was employed as loss function [37] and ADAM (Adaptive Moment Estimation) as optimizer [48]. Training was performed using a batch size of 64 along 470 iterations, implementing the cycling learning rate policy to obtain optimal classification results with fewer iterations [49]. Whereas different CNNs might require different number of iterations during training, we fixed this value because we wanted to select the best CNN architecture for the automatic classification of the different leukaemia types considered in this work. Moreover, the number of convolutional blocks to fine-tune was considered an hyperparameter to be selected. Several tests were performed varying the number of convolutional blocks to be trained: last block, last two, last three, last four and the entire network.

To select the model, CNN frameworks were trained and evaluated using two different approaches: 5-fold cross validation and hold-out. For the first case, we performed a random split of all the images of the whole dataset [16,450 single-cell images) into five equal subsets, ensuring that all the data from the same patient's smear was kept in the same subset. The same smear did not appear in two different folds. Moreover, folds were approximately balanced in the sense that the number of distinct smears was approximately the same in each fold. The mean and the standard deviation of the testing accuracy were computed among the five folds. For the hold-out approach, the *training set* was split into *train* and *validation sets* with 80% and 20% of the images, respectively. After training, the best CNN models were further evaluated using the *testing set*. The overall accuracy was selected as the main target performance parameter, so that we selected the networks that showed the highest value.

All the experiments were performed using PyTorch software libraries [37] and a Nvidia Titan XP GPU.

## 3. Results

### *3.1. First CNN classifier*

Table 3 shows the overall classification accuracies of the four evaluated CNNs with 5-fold cross validation and hold-out when changing the number of convolutional blocks trained with 470 iterations. Based on these results, the highest accuracies were obtained when fine-tuning the entire models. This means to train all the convolutional blocks with our own image dataset. Moreover, the highest testing accuracies were achieved with VGG16 and SENet154 architectures (above 85% using hold-out and above 86.8% with 5-fold cross validation) as it is shown in Table 3.

VGG16 showed an accuracy value of 88.4% using hold-out with respect to the best accuracy (87.2%) obtained with SENet154. Finally, we decided to select VGG16 together with the hold-out approach as the best model because of the following: 1) less overfitting was obtained as VGG16 showed better performance when making predictions with the new images of the *testing set*; and 2) it is a simpler architecture compared to the other CNN frameworks and had the best training and classification time, which is an advantage for a potential real-time implementation (see Table 3).

Table 4 shows the confusion matrix of the classification results of the six cell groups with the VGG16. The true positive rates (main diagonal) obtained for myeloid blasts (87%) and B-lymphoblasts (75%) were concluded not to be high enough for our purpose of building a tool to predict the diagnostic orientation of acute leukaemia in a clinical setting.

### *3.2. Sequential CNN classification system: ALNet*

In order to improve the true positive rates obtained for the recognition of myeloid blasts and B-lymphoblasts presented above, we proposed a new strategy with a two-step classification scheme (see Fig. 3), where two separate classifiers work in series: ALNet. The first step (module 1) consists of a VGG16 trained to distinguish abnormal promyelocytes among lymphocytes, monocytes, reactive lymphocytes and blast cells, which encompass both myeloid blasts and B-lymphoblasts. The second step (module 2) required a VGG with an increased number of convolutional layers (VGG19) to discriminate between myeloid blasts and B-lymphoblasts.

Although good results were obtained with VGG16 (testing accuracy of 85.7%), we tried to increase sensitivity for these two last

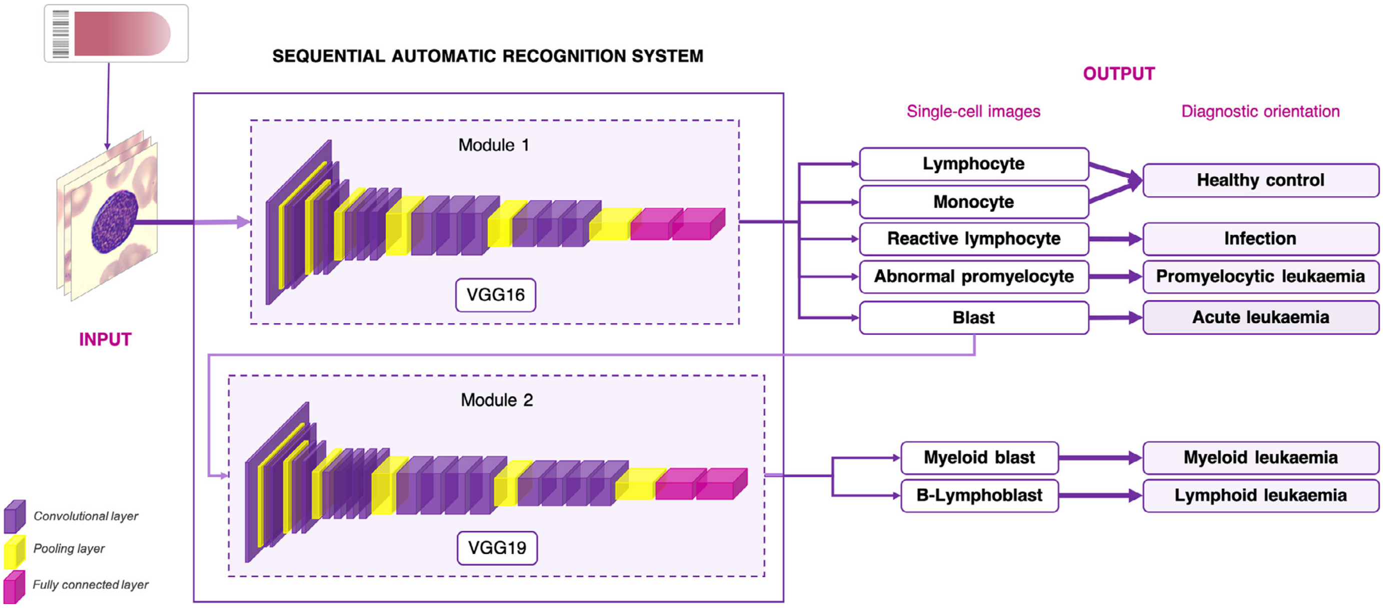

**Fig. 3.** Diagram illustrating the proposed sequential CNN-based system (ALNet) for the automatic recognition of acute leukaemia lineage. It starts with the selection of the cell images of an individual smear by the clinical pathologist, which are the inputs to the system. ALNet includes two consecutive modules. The first module recognises acute promyelocytic leukaemia (APL) and acute leukaemia (non-APL) from the remaining groups. Only the smears corresponding to patients with non-APL leukaemia (myeloid or lymphoid) are used for lineage classification through the second module.

**Table 5**
Overall training and testing accuracies obtained for module 2. The last two columns show the training time (in minutes) achieved when fine-tuning the entire models, and the classification speed (in seconds) when evaluating both models with the images of the *testing set*.

| | Accuracy | | Training time (min) | Class. speed (s) |
|---|---|---|---|---|
| Model | Training | Testing | | |
| **VGG16** | 95.4% | 85.7% | 2.42 | 8 |
| **VGG19** | 99.4% | 89.5% | 2.64 | 7 |

cell groups by using VGG19, which was the same architecture but with 16 convolutional layers instead of the 13 of VGG16. Regarding the design of the fully connected layers for the VGG19, 512 feature maps of 14 × 14 pixels were obtained from its convolutional blocks and converted to an array of 100,325 features, which were the input to the two fully connected layers after an average pooling. The first layer had 1,024 nodes and the second 512 nodes. The third fully connected layer consisted of two nodes, one node for each cell class (see Fig. 3). This increase of layers by using a VGG19 resulted into higher accuracies for both training and testing (99.4% and 89.5%, respectively) without compromising training time and classification speed (see Table 5).

Both VGGs were trained as explained in Section 2.3, using the images arranged as shown in Table 6 and performing fine-tuning to the whole convolutional blocks using hold-out. The first module exhibited a validation accuracy of 99.6% during training, while that of the second module was 99.4%.

The first assessment of ALNet was done through a blind classification of all the single-cell images of the *testing set* (see Table 6). We calculated the sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), precision or positive predictive value (PPV) and overall accuracy as follows:

$$Sensitivity\ (TPR) = \frac{TP}{TP+FN}$$

$$Specificity\ (TNR) = \frac{TN}{TN+FP}$$

$$Precision\ (PPV) = \frac{TP}{TP+FP}$$

$$Overall\ accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Table 7A shows the confusion matrix that summarizes the classification results for module 1. The true positive rates shown in the main diagonal were: 99.9% for lymphocytes, 97.6% for monocytes, 97.2% for reactive lymphocytes, 95.3% for abnormal promyelocytes and 91.7% for blasts (myeloid blasts and B-lymphoblasts). The overall classification accuracy was 94.2%. Table 7B summarizes the classification results for blast cells. As seen in the main diagonal, 99.4% of images corresponding to myeloid blasts and 82% to B-lymphoblasts were correctly classified. The overall classification accuracy was 89.5%. Specificity and precision values for each cell class of both modules are shown in Table 8.

### 3.3. Classification system ALNet in a clinical setting

Going towards the final objective of this work, we assessed ALNet to predict patient's diagnosis using the smear as a classification unit, trying to emulate the way pathologists interpret results in clinical laboratories. The system input was a set of cell images of an individual smear selected by the clinical pathologist and the output was the prediction of one of the following diagnoses: APL, AML, ALL or infection (see Fig. 3). The diagnosis was predicted by identifying the cell class that predominated in the smear. A threshold was determined such that the diagnosis will be predicted by the cell class with the percentage of images classified above this value. We used all the smears of the *training set* to perform a multiclass Receiver Operating Characteristics (ROC) analysis with the statistical R software [50]. Results are shown in Fig. 4. It was found that 50% of the cell images correctly classified was the best threshold to predict patient's initial diagnosis through the smear, obtaining values of 1 for the area under the curve. Such satisfactory result could be explained because only 91 out of 14,985 (0.6%) single-cell images of the *training set* were misclassified, and

**Table 6**
Distribution of the cell images and smears used to train and test the sequential classification system with two modules (ALNet). Images and smears are grouped by class for each dataset and the number of healthy controls and patients is provided.

| | | Number of images | | | Number of smears | | Number of controls and patients |
|---|---|---|---|---|---|---|---|
| | | Training | | Testing | Training | Testing | |
| | Cell type | Train | Validation | | | | |
| **Module 1** | **Lymphocytes** | 2,510 | 778 | 312 | 82 | 13 | 100 |
| | **Monocytes** | 945* | 260 | 117 | 83 | 4 | |
| | **Reactive lymphocytes** | 1,462* | 390 | 459 | 168 | 46 | 191 |
| | **Abnormal promyelocytes** | 2,035* | 540 | 1,358 | 19 | 14 | 15 |
| | **Blasts** | 2,790 | 697 | 1,797 | 272 | 57 | 133 |
| **Module 2** | **Myeloid blasts** | 2,296+ | 574 | 774 | 224 | 30 | 96 |
| | **B-Lymphoblasts** | 494+ | 123 | 1,023 | 48 | 27 | 37 |

* Cell groups up-sampled to 2,500 images with data augmenting techniques
+ Cell groups up-sampled to 2,900 images with data augmenting techniques

**Table 7**
Confusion matrix of the classification results (in %) for the images of the *testing set* of modules 1 (A) and 2 (B). Rows indicate the true class and columns represent the predicted class supplied by the network. Diagonal values are the true positive rates for each cell type. The overall classification accuracies of the first (A) and second module (B) were 94.2% and 89.5%, respectively.

| A | | Predicted class | | | | |
|---|---|---|---|---|---|---|
| | | Lymphocytes | Monocytes | Reactive lymphocytes | Abnormal promyelocytes | Blasts |
| True class | **Lymphocytes** | **99.9** | 1 | 0 | 0 | 0 |
| | **Monocytes** | 0.8 | **97.6** | 0 | 0.8 | 0.8 |
| | **Reactive lymphocytes** | 0.2 | 0.2 | **97.2** | 0 | 2.4 |
| | **Abnormal promyelocytes** | 0 | 0.4 | 0 | **95.3** | 4.3 |
| | **Blasts** | 2.5 | 0.2 | 1.3 | 4.3 | **91.7** |

| **B** | | Predicted class | |
|---|---|---|---|
| | | **Myeloid blasts** | **B-Lymphoblasts** |
| True class | **Myeloid blasts** | **99.4** | 0.6 |
| | **B-Lymphoblasts** | 18 | **82** |

**Table 8**
Sensitivity, specificity and precision values of module 1 and 2 of ALNet regarding the classification results of individual cell images.

| | Module 1 | | | | | Module 2 | |
|---|---|---|---|---|---|---|---|
| | Lymphocytes | Monocytes | Reactive lymphocytes | Abnormal promyelocytes | Blasts | Myeloid blasts | B-Lymphoblasts |
| **Sensitivity** | 99.9% | 97.6% | 97.2% | 95.3% | 91.7% | 99.4% | 82.0% |
| **Specificity** | 98.7% | 99.7% | 99.3% | 97.1% | 96.9% | 82.0% | 99.4% |
| **Precision** | 86.8% | 90.5% | 94.9% | 94.2% | 95.9% | 80.7% | 99.4% |

this number was nearly imperceptible when focusing on the whole smear as classification unit.

Such threshold may be interpreted in such a way that if more than 50% of the cell images of a smear are classified as myeloid blasts, for example, the predicted diagnosis of the patient to whom this smear belongs is acute myeloid leukaemia. Whereas if similar percentages are obtained for more than one class, in this case it is not possible to predict a diagnosis and the system considers the smear as belonging to an "unknown" diagnostic group.

Once the threshold was determined, the automatic recognition system was assessed using one by one all the smears in the *testing set*, which were not previously used (see Table 6). Confusion matrix in Table 9A shows the classification results of the first module. Sensitivity, specificity and precision values of 100% were obtained for all the categories of the first module. In consequence, using the first classification module of ALNet, we correctly detected the following groups: 1) healthy controls, 2) patients with infection, 3) patients with APL and 4) patients with acute leukaemia non-APL.

The smears corresponding to patients with non-APL leukaemia (AML or ALL) were classified through the second module, and the results are shown in the confusion matrix in Table 9B. Regarding AML, sensitivity, specificity and precision values of 100%, 92.3% and 93.7%, respectively, were obtained. As for ALL, a sensitivity of 89% and specificity and precision values of 100% were obtained. The overall accuracy of individual smears was 94.7% (see Table 9B).

At the end of the two-step classification with ALNet, we obtained the correct diagnostic prediction for all patients with APL and AML. With respect to the ALL group, a total of 24 from 27 smears were correctly classified, being the remaining three recognized as AML (two smears) or unknown (one smear).

In addition, to further evaluate the ALNet performance in a clinical setting, we used a new image dataset from two other hospitals (*Josep Trueta* and *Germans Trias i Pujol*). This dataset contained a total of 381 images of blast cells (322 myeloid blasts and 59 B-lymphoblasts) stained with May Grünwald-Giemsa and acquired by the CellaVision analyser used in each laboratory. Fig. 5 shows representative examples of myeloid blast and B-lymphoblast images from the three datasets used to evaluate ALNet.

When classifying by single-cell images, 98% of blast cells were correctly classified by module 1. Module 2 correctly classified 82% of myeloid blasts and 64% of B-lymphoblasts. When using the threshold of 50% to predict the diagnosis from the smear, in the first classification module of ALNet all the smears (100%) were classified as acute leukaemia non-APL. Regarding the leukaemia

**Table 9**
Confusion matrix of the classification results (in %) for the smears of the *testing set* of modules 1 (A) and 2 (B) taking the threshold of 50% into consideration. Rows indicate the true diagnosis and columns represent the predicted diagnosis supplied by the classifier. Diagonal values are the true positive rates for each smear (in brackets the number of smears). The overall classification accuracies of the first (A) and second module (B) were 100% and 94.7%, respectively. LY, lymphocytes; MO, monocytes; APL, acute promyelocytic leukaemia; AL, acute leukaemia; AML, acute myeloid leukaemia; ALL, acute lymphoid leukaemia; UNK, unknown.

| A | | Predicted diagnosis | | | | |
|---|---|---|---|---|---|---|
| | | Control (LY) | Control (MO) | Infection | APL | AL non-APL |
| True diagnosis | **Control (LY)** | **100% (13)** | 0 | 0 | 0 | 0 |
| | **Control (MO)** | 0 | **100% (4)** | 0 | 0 | 0 |
| | **Infection** | 0 | 0 | **100% (46)** | 0 | 0 |
| | **APL** | 0 | 0 | 0 | **100% (14)** | 0 |
| | **AL non-APL** | 0 | 0 | 0 | 0 | **100% (57)** |

| **B** | | Predicted diagnosis | | |
|---|---|---|---|---|
| | | **AML** | **ALL** | **UNK** |
| True diagnosis | **AML** | **100% (30)** | 0 | 0 |
| | **ALL** | 7% (2) | **89% (24)** | 4% (1) |

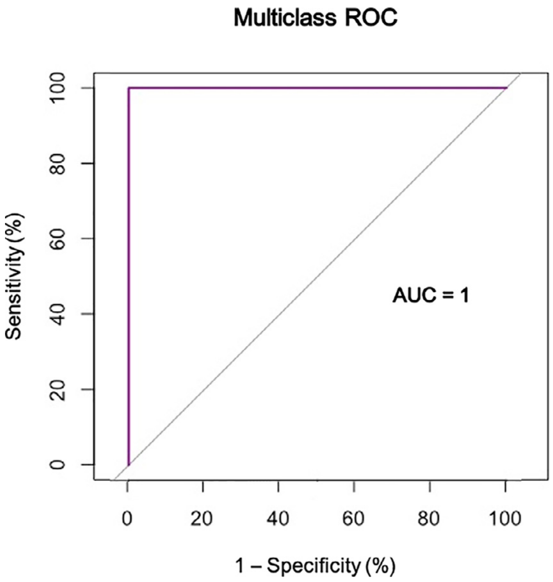

**Fig. 4.** Plot of the multiclass ROC analysis to obtain the best threshold to predict the diagnosis of the subtype of acute leukaemia. It was obtained by averaging the single ROC curves obtained for each of the classes under study. The best threshold is defined as the minimum percentage of images required in a cell class to predict a diagnosis. The 50% resulted to be the best threshold, obtaining an AUC of 1. AUC, area under the curve; ROC, Receiver Operating Characteristic.

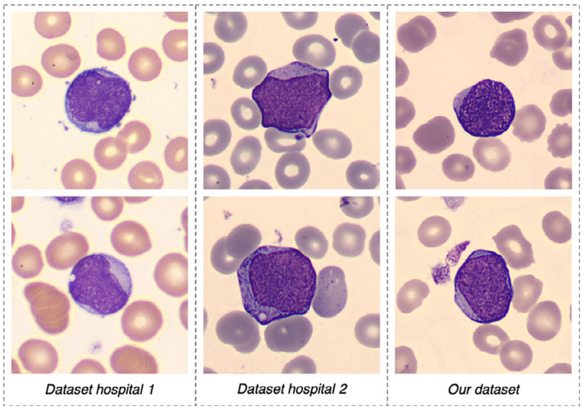

**Fig. 5.** Examples of B-lymphoblast images (first row) and myeloid blast images (second row) of the three datasets included in this study to evaluate the performance of ALNet.

lineage, 3/5 smears corresponding to AML (60%) and 1/2 ALL smears (50%) were correctly classified.

## 4. Discussion

For the first diagnostic orientation of acute leukaemia in clinical laboratories, it is important the detection of quantitative abnormalities in leucocyte count, haemoglobin level or thrombocyte count, which triggers the smear review. The detection of blasts circulating in blood is the subsequent step. Thus, cell morphology is crucial for the initial diagnosis of leukaemia lineage to apply the suitable treatment and avoid delays in medical procedures, primarily in acute promyelocytic leukaemia. This is why the automatic recognition of blasts has been a target for automated solutions, tools and methods within the artificial intelligence framework [6,7].

Recently, previous studies described deep learning approaches to recognise lymphoblasts from normal leukocytes [19,22–24,26,27]. Table 10 shows a comparison of the overall accuracy obtained with our proposed system with that obtained in other works in the literature. It also summarizes the number of images and the CNN frameworks used by these previous authors. These studies achieved accuracy values above 88% for lymphoblasts, but it is fair to mention that the significant morphological differences with normal leukocytes make the classification problem more accessible. However, when other authors tried to distinguish B-lymphoblasts from mononuclear cells such as lymphocytes, the sensitivity for B-lymphoblasts decreased to 81.5% [21]. Many of these studies focused on a ‘binary’ classification (disease vs. normal), which is not a realistic approach and does not reflect the real-life complexity of haematological malignancies diagnosis [7]. Moreover, those studies which classified lymphoblasts subtypes [20,21] were based on the FAB classification, which currently has been replaced by the WHO 2016 classification [36] for clinical practise. The sensitivity for B-lymphoblasts (82%) obtained in this work (Table 7B) is the best result achieved nowadays in the automatic classification among blasts using deep learning techniques. Moreover, the high sensitivity (95.3%) obtained with abnormal promyelocytes (Table 7A) when differentiating them from other blasts overcomes the accuracy of 41% (Table 10) obtained by previous authors [25] when differentiating abnormal promyelocytes from myeloid blasts. Besides this, nearly none false positives were obtained when discriminating normal mononuclear and reactive cells from blasts (as seen in Table 7A).

In our work, the challenge was twofold: 1) differentiate blasts among other mononuclear cells, and 2) discern between myeloid

**Table 10**
Comparison of the overall accuracy obtained with the proposed system (ALNet) with that of other works in the literature. CNN, convolutional neural network; fc, fully connected layers; NR, not reported; LDP, local directional pattern; SCA, sine cosine algorithm; BL, blast cells; Ab. prom, abnormal promyelocytes; SVM, support vector machine.

| Work | Original N° images | Feature extraction | Classification | Accuracy Leukaemia detection | Accuracy Subtype of leukaemia |
|---|---|---|---|---|---|
| Thahn *et al.*(19) | 108 | CNN | fc | 96.6% | NR |
| Shafique *et al.*(20) | 260 | CNN (AlexNet) | fc | 99.5% | 96.06% |
| Pansombut *et al.*(21) | 363 | CNN | fc | 81.74% | 81.5% B-lymphoblasts<br>68.9% T-lymphoblasts |
| Ahmed *et al.*(22) | 354 | CNN | fc | 88.25% | NR |
| Jha *et al.*(23) | 260 | LDP | SCA - fc | 98.7% | NR |
| Prellberg *et al.*(24) | 12,528 | CNN (ResNeXt50) | fc | 88.91% | NR |
| Matek *et al.*(25) | 18,365 (3,312 BL) | CNN (ResNeXt) | fc | 90% | 94% myeloblasts 41% Ab. prom |
| Loey *et al.*(26) | 564 | CNN (AlexNet) | fc | 100% | NR |
| Vogado *et al.*(27) | 377 | CNN (AlexNet, CaffeNet, VGG) | SVM | 99% | NR |
| Di Ruberto *et al.*(28) | 33 | CNN (AlexNet) | SVM | 94.1% | NR |
| Rehman *et al.*(29) | 330[+] | CNN (AlexNet) | fc | 97.78% | NR |
| Huang *et al.*(30) | 1,322[+] | CNN (DenseNet121) | fc | 95.3% | 95.25% |
| **Proposed system (ALNet)** | **16,450 (4,825 BL)** | **CNN (VGG)** | **fc** | **94.2% (cell) 100% (smear)** | **89.5% (cell) 94.7% (smear)** |

[+] Datasets of bone marrow digital images

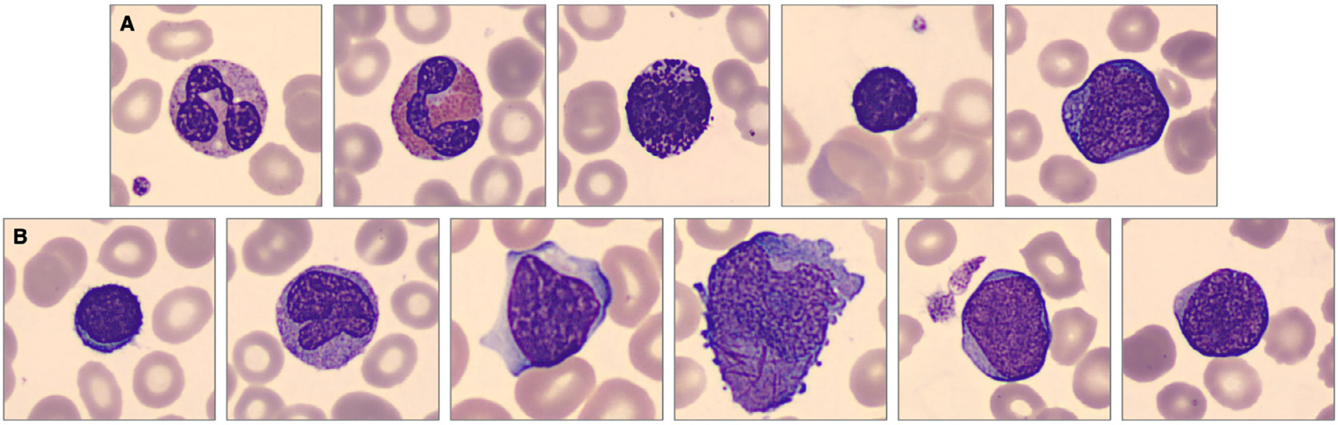

**Fig. 6.** (A) Examples of cell types included in previous studies from the literature for leukaemia classification (from left to right): neutrophil, eosinophil, basophil, lymphocyte and B-lymphoblast. (B) Cell groups included in this study (from left to right): lymphocyte, monocyte, reactive lymphocyte, abnormal promyelocyte, myeloblast and B-lymphoblast.

blasts and B-lymphoblasts because of the overlapping morphological characteristics that they exhibit. It may be debatable why this work did not include neutrophils or eosinophils or erythroblasts in the classification problem when they may be present in the smear. As shown in Fig. 6A, lymphoblasts exhibit very different morphology with respect to mature leukocytes, such as neutrophils, eosinophils and basophils since their nucleus is lobulated and their cytoplasm shows abundant granules. Fig. 6B shows representative examples of the groups addressed in this study (from left to right: lymphocyte, monocyte, reactive lymphocyte, abnormal promyelocyte, myeloblast and B-lymphoblast). By comparing the cell groups in Fig. 6A (from left to right: neutrophil, eosinophil, basophil, lymphocyte and B-lymphoblast) with the groups in Fig. 6B, it is visible that the automatic differentiation in the second case is more complex.

In a previous work [16] using CNNs, our group achieved high performance with a fine-tuned VGG16 for the recognition of eight groups of normal cells circulating in peripheral blood. We focused on normal leukocytes, obtaining accuracies for neutrophils, eosinophils, immature granulocytes (metamyelocytes, myelocytes and promyelocytes) and erythroblasts of 99.6%, 99.6%, 92.8% and 91.8%, respectively, among the other cell groups. The current paper was focused on the differentiation of the lineage of blast cells with the aim of designing a model to assist the clinical pathologist. Computational assistance is more needed for blast lineage discrimination than for blast versus normal leukocyte differentiation. The integration of the first classifier [16] and the current ALNet is a next step which can be taken in the near future.

In a previous publication by the authors [35], the objectives of this paper were addressed using traditional machine learning algorithms, where a single model based on linear discriminant analysis achieved the highest classification accuracy. In the present work, our purpose was to improve the previous accuracy for the leukaemia lineage differentiation, which was 85.8%. The new strategy herein was to design two separated classifiers working sequentially. The second classifier was specialized to face the most difficult recognition task, which is the automatic differentiation between myeloid blasts and B-lymphoblasts. Using this strategy, sensitivity values were increased to 95.3% for abnormal promyelocytes, 99.4% for myeloid blasts and 82% for B-lymphoblasts (see Table 11). Furthermore, traditional approaches required to segment and extract features manually. Indeed, segmentation presents many challenges because of the complex cell morphology, the variability of the blood smear staining and the general conditions of image acquisition. All these aspects may result in a decrease of the classification accuracy of these traditional models. Table 12 shows a com-

**Table 11**
True positive rates for each cell type, comparing the proposed deep learning system (ALNet) with the approach previously published by our group. In bold are shown the improved sensitivity values achieved with ALNet.

| | True positive rates | |
|---|---|---|
| Type of cell | Boldú*et al.* [35] | Proposed system |
| Lymphocytes | 97.6% | **99.9%** |
| Monocytes | 93.0% | **97.6%** |
| Reactive lymphocytes | 97.7% | 97.2% |
| Abnormal promyelocytes | 72.6% | **95.3%** |
| Myeloblasts | 80.8% | **99.4%** |
| B-Lymphoblasts | 78.9% | **82.0%** |

**Table 12**
Performance comparison between traditional machine learning approaches previously published by our group and the deep learning models proposed in this study. LDA, linear discriminant analysis; SVM, support vector machine; RF, random forest; KNN, k-nearest neighbours; Bayes, naïve Bayes.

| | Method | Accuracy |
|---|---|---|
| **Machine learning approaches** (35) | **LDA** | 85.8% |
| | **SVM** | 83.5% |
| | **RF** | 75.4% |
| | **KNN** | 74.8% |
| | **Bayes** | 68.3% |
| **Deep learning approaches** | **VGG16** | 94.6% |
| | **ResNet101** | 93.3% |
| | **DenseNet121** | 93.6% |
| | **SENet154** | 94.6% |

parison of the accuracies when using those models and the CNN models investigated herein, which are higher in accordance to the results reported by previous authors [21,22,26].

The results of this study demonstrate that the best classification of blast cells is achieved with the proposed sequential system, being VGG the best architecture for both modules. This is consistent with the most frequently used networks in the literature, where simple architectures predominate to address the recognition of acute leukaemia [20,27,29]. Moreover, we concluded from our experience that using hold-out as validation technique with 2,500 images per group made possible obtaining satisfactorily high accuracies as similar accuracies were obtained with the 5-fold cross validation approach. This was consistent with our previous publication [16] and results published from previous authors [26].

An important aspect to remark in this work is the number of images involved and their high quality, being the largest dataset used for leukaemia classification as it is illustrated in Table 10. We used a dataset with over 16,000 peripheral blood cell images, 4,825 being blast cell images, in comparison with the 3,312 blast cell images used by Matek *et al.* [25]. Furthermore, we guarantee the confirmation of their labels through other complementary tests (ground truth). Not only high-quality and a large number of images are required for developing diagnostic systems, but also the availability of images properly annotated by experts, which is scarce in literature [7,8,19]. The quality of a dataset is essential to observe morphological characteristics which can lead towards a diagnosis, not only for daily clinical practise but also to obtain robust models avoiding overfitting [16].

Visualizing the feature maps generated by the intermediate convolutional layers of a CNN framework could give us an idea of which patterns the network extracts, and thus help to interpret classification results. With this purpose, both our ALNet system and the original VGG16 ImageNet model were fed with two representative examples of cell images. Afterwards, we extracted the feature maps from the same convolutional layers to compare them, as it is illustrated in Fig. 7.

Fig. 7A shows original images of a myeloid blast and an abnormal promyelocyte. Fig. 7B displays the feature map 202 of 256 from the convolutional block 3, layer 1. In the feature maps generated by the original VGG16 (left) we can barely notice the cell nucleus, and in the case of the abnormal promyelocyte neither the cell outline. At this depth the original network cannot detect diagonal and horizontal lines to differentiate parts of the cells of the input images. With respect to the feature maps obtained from the VGG16 of ALNet (right), our model is able to precisely detect the cell outline, nucleus shape and texture, along with the recognition of the red blood cells and their central pallor and the platelet next to the myeloid blast. The histogram of the myeloid blast (Fig. 7C) presents a narrow and flatter peak with the pixels very localized within a short intensity range between 100 and 150, and also having a greater number of darker pixels mostly localized in the cell nucleus. In contrast, the histogram of the abnormal promyelocyte shows a peak with lower height, which covers many pixels with different intensity levels. This indicates more information variety due to the presence of the intense azurophil granulation and splinters on its larger cytoplasm.

Fig. 7D shows another case where the improvement of ALNet in detecting relevant characteristics of the cells is visually interpretable. At this stage ALNet is capable of detecting the bilobed nucleus of the abnormal promyelocyte and the nucleolus of the myeloid blast, while the feature map from the original network does not capture almost any information from the images. It is important to mention that subtle morphological differences in the nucleus shape, chromatin texture and cytoplasmic granulation can make the difference among the cell types included in this study.

When evaluating ALNet with images acquired in two other hospitals, sensitivity for the leukaemia lineage discrimination decreased. Previous authors [22] reported that the overall accuracy of their CNN model decreased 6.51% when it was evaluated with a different image dataset. Although all datasets of this work were stained with the ‘gold standard’ May Grünwald-Giemsa technique used in clinical laboratories, we observed that variations in the optical conditions and resolution of microscopic images affected the classification accuracy of the CNN models (see Fig. 5). It is known that colour information is involved in the quantification of cytoplasmic basophilia and granulation, both being very important characteristics in the classification of myeloid blasts and lymphoblasts [51]. This is why changes related to image clarity, colour scale or resolution could mislead their differentiation.

When developing new diagnostic support tools for laboratory practise, it is important to consider individual patients when organising datasets for training and assessing the system. Using images from the same smear for both training and assessment could cause an accuracy overestimation, what happens in almost all the studies in the literature. To avoid this, we followed the procedure established in [35], by splitting the initial dataset into two sets of different smears. An excellent diagnostic prediction was achieved as the system differentiated all normal smears from those related to infections and with respect to smears containing blast cells (as seen in Table 9A). This satisfactory performance was also obtained with smears corresponding to patients with acute leukaemia from two other hospitals. Moreover, the system presented a very high sensitivity (100%) for the detection of myeloid leukaemia, and high specificity and precision (100%) for promyelocytic and lymphoid leukaemia for the smears of our own dataset.

These results showed that the approach proposed in this work could be suitable for clinical laboratory practise. A current limitation for the practical implementation of ALNet in other laboratories may be that a single-institution data source has been used to train the model, which could result in accuracy variations related to the image staining procedure, as well as optical and resolution of microscopic images. To deal with this, our group has some work in

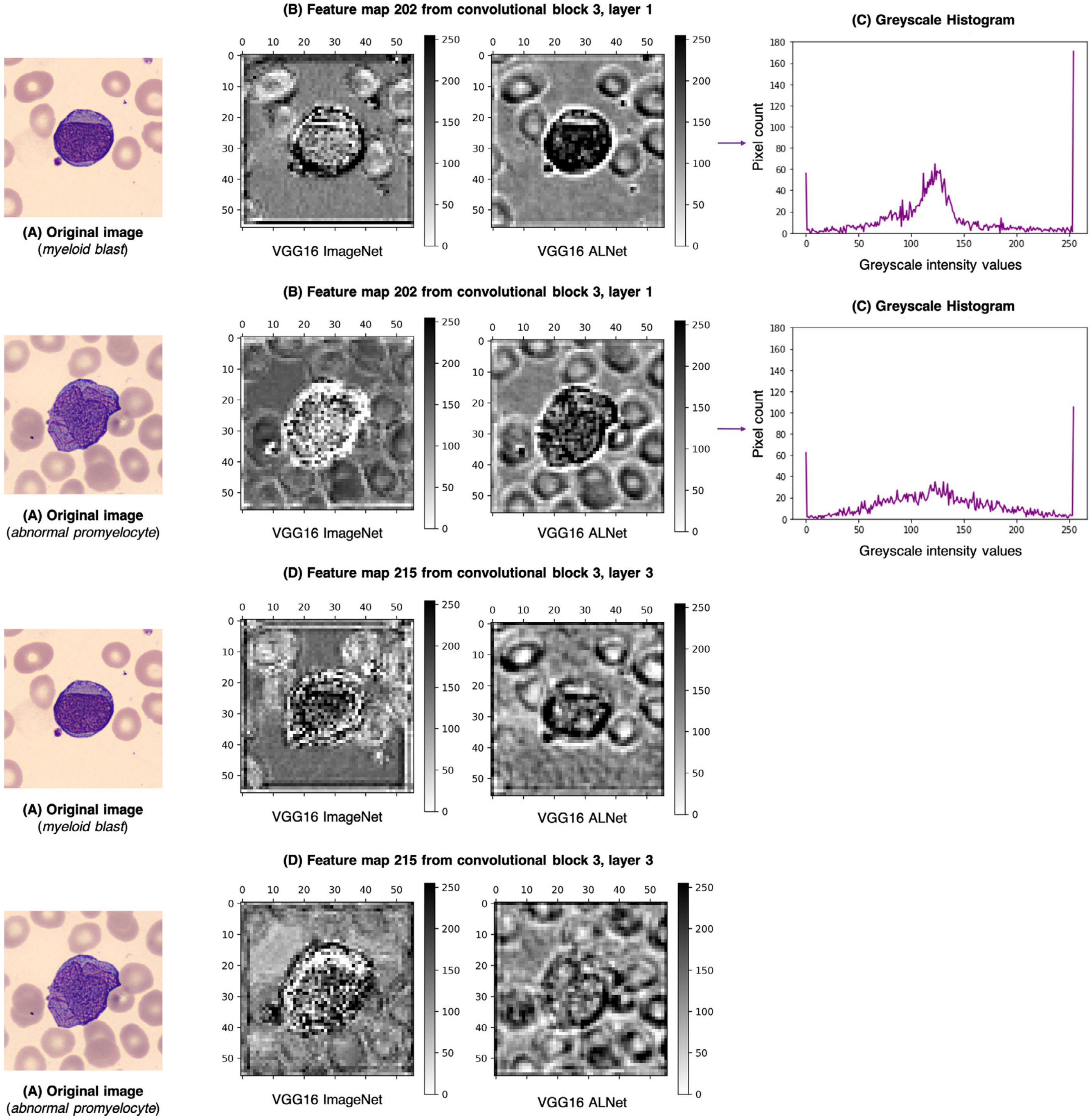

**Fig. 7.** (A) Original images of a myeloid blast from AML and an abnormal promyelocyte from APL. (B and D) Feature maps from VGG16 ImageNet weights (left) compared with feature maps from VGG16 of ALNet (right) for the classification of five classes of peripheral blood cells. (C) Greyscale histograms calculated from the features maps obtained from the VGG16 of ALNet. AML, acute myeloid leukaemia; APL, acute promyelocytic leukaemia.

progress using new models based on Generative Adversarial Networks (GANs) to standardize the images that feed the CNNs.

## 5. Conclusions

The final contribution of this paper is a predictive model designed with two serially connected convolutional networks and trained using a dataset with over 16,000 blood cell images obtained from clinical practise. It is proposed to assist clinical pathologists in the diagnosis of acute leukaemia during the blood smear review. It has been proved to distinguish neoplastic (leukaemia) and non-neoplastic (infections) diseases, as well as recognise the leukaemia lineage.

## Authors' contributions

Laura Boldú designed the datasets, developed the classifiers, performed and evaluated the experiments, reviewed the literature and contributed to the manuscript writing.

Anna Merino supervised the overall project and contributed to the manuscript writing and editing.

Andrea Acevedo contributed to the algorithmic implementations and analysis of classification results.

Angel Molina contributed to data collection and morphological annotation.

José Rodellar advised on the deep learning classification models and contributed to the manuscript writing and editing.

## Declaration of Competing Interest

We wish to confirm that there are not known conflicts of interest associated with this publication and that this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors that could have influenced its outcome.

## Acknowledgments

This work is part of a research project funded by the Ministry of Science and Innovation of Spain, with reference PID2019-104087RB-I00.

We thank Maite Serrando and Cristian López-Indiano for sharing the images acquired in their hospital.

Laura Boldú thanks financial support for the scholarship received from Siemens Healthcare S.L.U as a part of a collaborative agreement with the Hospital Clinic of Barcelona.

## References

[1] A Miranda-Filho, M Piñeros, J Ferlay, I Soerjomataram, A Monnereau, F. Bray, Epidemiological patterns of leukaemia in 184 countries: a population-based study, Lancet Haematol. 5 (1) (2018) e14–e24.

[2] A Merino, L Boldú, A. Ermens, Acute myeloid leukaemia: How to combine multiple tools, Int. J. Lab. Hematol. 00 (2018) 1–11 Available from, doi:10.1111/ijlh.12831.

[3] J Rodellar, S Alférez, A Acevedo, A Molina, A. Merino, Image processing and machine learning in the morphological analysis of blood cells, Int. J. Lab. Hematol. (40) (2018) 46–53.

[4] S Alférez, A Merino, L Bigorra, L Mujica, M Ruiz, J. Rodellar, Automatic recognition of atypical lymphoid cells from peripheral blood by digital image analysis, Am. J. Clin. Pathol. 143 (2015) 168–176.

[5] C Briggs, I Longair, M Slavik, K Thwaite, R Mills, V Thavaraja, et al., Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system, Int. J. Lab. Hematol. 31 (1) (2009) 48–60.

[6] MA Alsalem, AA Zaidan, BB Zaidan, M Hashim, HT Madhloom, ND Azeez, et al., A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations, Comput. Methods Programs Biomed. 158 (2018) 93–112.

[7] HT Salah, IN Muhsen, ME Salama, T Owaidah, SK. Hashmi, Machine learning applications in the diagnosis of leukemia: current trends and future directions, Int. J. Lab. Hematol. 41 (6) (2019) 717–725.

[8] H El Achi, JD. Khoury, Artificial Intelligence and digital microscopy applications in diagnostic hematopathology, Cancers (Basel) 12 (4) (2020) 797.

[9] N Radakovich, M Nagy, A. Nazha, Machine learning in haematological malignancies, Lancet Haematol. 7 (7) (2020) e541–e550.

[10] R Shouval, JA Fein, B Savani, M Mohty, A. Nagler, Machine learning and artificial intelligence in haematology, Br. J. Haematol. (2020).

[11] M Habibzadeh, A Krzyżak, T. Fevens, White blood cell differential counts using convolutional neural networks for low resolution images, Int. Conf. Artif. Intell. Soft Comput. (2013) 263–274.

[12] M-C Su, C-Y Cheng, P-C. Wang, A neural-network-based approach to white blood cell classification, Sci. world J. (2014).

[13] J Rawat, A Singh, HS Bhadauria, J Virmani, JS. Devgun, Application of ensemble artificial neural network for the classification of white blood cells using microscopic blood images, Int. J. Comput. Syst. Eng. 4 (2–3) (2018) 202–216.

[14] F Qin, N Gao, Y Peng, Z Wu, S Shen, A. Grudtsin, Fine-grained leukocyte classification with deep residual learning for microscopic images, Comput. Methods Programs Biomed. 162 (2018) 243–252.

[15] AI Shahin, Y Guo, KM Amin, AA. Sharawi, White blood cells identification system based on convolutional deep neural learning networks, Comput. Methods Programs Biomed. 168 (2019) 69–80.

[16] A Acevedo, S Alférez, A Merino, L Puigví, J. Rodellar, Recognition of peripheral blood cell images using convolutional neural networks, Comput. Methods Programs Biomed. 180 (2019) 105020.

[17] JW Choi, Y Ku, BW Yoo, J-A Kim, DS Lee, YJ Chai, et al., White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks, PLoS One 12 (12) (2017) e0189259.

[18] H El Achi, T Belousova, L Chen, A Wahed, I Wang, Z Hu, et al., Automated diagnosis of lymphoma with digital pathology images using deep learning, Ann. Clin. Lab. Sci. 49 (2) (2019) 153–160.

[19] TTP Thanh, C Vununu, S Atoev, S-H Lee, K-R. Kwon, Leukemia blood cell image classification using convolutional neural network, Int. J. Comput. Theory Eng. 10 (2) (2018) 54–58.

[20] S Shafique, S. Tehsin, Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks, Technol. Cancer Res. Treat. 17 (2018) 1533033818802789.

[21] T Pansombut, S Wikaisuksakul, K Khongkraphan, Phon-on A. Convolutional neural networks for recognition of lymphoblast cell images, Comput. Intell. Neurosci. (2019).

[22] N Ahmed, A Yigit, Z Isik, A. Alpkocak, Identification of leukemia subtypes from microscopic images using convolutional neural network, Diagnostics 9 (3) (2019) 104.

[23] KK Jha, HS. Dutta, Mutual information based hybrid model and deep learning for Acute Lymphocytic Leukemia detection in single cell blood smear images, Comput. Methods Programs Biomed. 179 (2019) 104987.

[24] J Prellberg, O. Kramer, Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks, in: ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging, Springer, 2019, pp. 53–61.

[25] C Matek, S Schwarz, K Spiekermann, C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks, Nat. Mach. Intell. 1 (11) (2019) 538–544.

[26] M Loey, M Naman, H. Zayed, Deep transfer learning in diagnosing leukemia in blood cells, Computers 9 (2) (2020) 29.

[27] LHS Vogado, RMS Veras, FHD Araujo, RR V Silva, KRT. Aires, Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification, Eng. Appl. Artif. Intell. 72 (2018) 415–422.

[28] C Di Ruberto, A Loddo, G. Puglisi, Blob detection and deep learning for leukemic blood image analysis, Appl. Sci. 10 (3) (2020) 1176.

[29] A Rehman, N Abbas, T Saba, SI UR Rahman, Z Mehmood, H. Kolivand, Classification of acute lymphoblastic leukemia using deep learning, Microsc. Res. Tech. 81 (11) (2018) 1310–1317.

[30] F Huang, P Guang, F Li, X Liu, W Zhang, W. Huang, AML, ALL, and CML classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network: a STARD compliant diagnosis research, Medicine (Baltimore) 99 (45) (2020) 1–8.

[31] G Gutiérrez, A Merino, A Domingo, JM Jou, JC. Reverter, EQAS for peripheral blood morphology in Spain: a 6-year experience, Int. J. Lab. Hematol. 30 (6) (2008) 460–466.

[32] D Chabot-Richards, K. Foucar, Does morphology matter in 2017? An approach to morphologic clues in non-neoplastic blood and bone marrow disorders, Int. J. Lab. Hematol. 39 (2017) 23–30.

[33] BJ. Bain, in: Leukaemia Diagnosis, 4th ed., Wiley-Blackwell, Chichester, UK, 2010, pp. 68–73.

[34] KA Breen, D Grimwade, BJ. Hunt, The pathogenesis and management of the coagulopathy of acute promyelocytic leukaemia, Br. J. Haematol. 156 (1) (2012) 24–36.

[35] L Boldú, A Merino, S Alférez, A Molina, A Acevedo, J. Rodellar, Automatic recognition of different types of acute leukaemia in peripheral blood by image analysis, J. Clin. Pathol. 72 (11) (2019) 755–761.

[36] DA Arber, A Orazi, R Hasserjian, J Thiele, MJ Borowitz, MM Le Beau, et al., The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia, Blood 127 (20) (2016) 2391–2405.

[37] V. Subramanian, Deep Learning with PyTorch: a Practical Approach to Building Neural Network Models Using PyTorch, Packt Publishing Ltd, 2018.

[38] G Litjens, T Kooi, BE Bejnordi, AAA Setio, F Ciompi, M Ghafoorian, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[39] J Yosinski, J Clune, Y Bengio, H. Lipson, How transferable are features in deep neural networks? in: Advances in Neural Information Processing Systems, 2014, pp. 3320–3328.

[40] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019.

[41] O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[42] K Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv Prepr arXiv14091556.

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

[44] G Huang, Z Liu, L Van Der Maaten, KQ. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[45] J Hu, L Shen, S Albanie, G Sun, E. Wu, Squeeze-and-Excitation Networks, 2017.

[46] S Alferez, A Merino, L Boldú, A Acevedo, A Molina, J. Rodellar, A deep learning approach to automatically classify pathological cell images in peripheral blood, in: ISLH 2019 Abstract Proceedings, 2019.

[47] A Merino, S Alférez, L Boldú, A Molina, L Puigví, A Acevedo, et al., Automatic differentiation of acute leukaemia, lymphoma and reactive lymphocytes in peripheral blood using a novel convolutional network, in: ISLH 2020 Abstract Proceedings, 2020.

[48] DP Kingma, Ba J. Adam, A Method for Stochastic Optimization, 2014 arXiv Prepr arXiv14126980.

[49] LN. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV, 2017, pp. 464–472.

[50] DJ Hand, RJ. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2) (2001) 171–186.

[51] A Merino, L Puigví, L Boldú, S Alférez, J. Rodellar, Optimizing morphology through blood cell image analysis, Int. J. Lab. Hematol. 40 (2018) 54–61.