Computer Methods and Programs in Biomedicine 240 (2023) 107629

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Automatic normalized digital color staining in the recognition of abnormal blood cells using generative adversarial networks☆

Kevin Barrera[a], José Rodellar[a], Santiago Alférez[a,*], Anna Merino[b]

[a] Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona East Engineering School, Department of Mathematics, Barcelona, Spain
[b] Hospital Clínic of Barcelona-IDIBAPS, Biochemistry and Molecular Genetics Department, CORE Laboratory, Biomedical Diagnostic, Barcelona, Spain

ARTICLE INFO

Article history:
Received 7 March 2023
Revised 23 May 2023
Accepted 28 May 2023

Keywords:
Deep learning
Blood cell morphology
Digital staining
Generative adversarial networks
Convolutional neural networks
Blood cell automatic recognition

ABSTRACT

Background and Objectives: Combining knowledge of clinical pathologists and deep learning models is a growing trend in morphological analysis of cells circulating in blood to add objectivity, accuracy, and speed in diagnosing hematological and non-hematological diseases. However, the variability in staining protocols across different laboratories can affect the color of images and performance of automatic recognition models. The objective of this work is to develop, train and evaluate a new system for the normalization of color staining of peripheral blood cell images, so that it transforms images from different centers to map the color staining of a reference center (RC) while preserving the structural morphological features.

Methods: The system has two modules, GAN1 and GAN2. GAN1 uses the PIX2PIX technique to fade original color images to an adaptive gray, while GAN2 transforms them into RGB normalized images. Both GANs have a similar structure, where the generator is a U-NET convolutional neural network with ResNet and the discriminator is a classifier with ResNet34 structure. Digitally stained images were evaluated using GAN metrics and histograms to assess the ability to modify color without altering cell morphology. The system was also evaluated as a pre-processing tool before cells undergo a classification process. For this purpose, a CNN classifier was designed for three classes: abnormal lymphocytes, blasts and reactive lymphocytes.

Results: Training of all GANs and the classifier was performed using RC images, while evaluations were conducted using images from four other centers. Classification tests were performed before and after applying the stain normalization system. The overall accuracy reached a similar value around 96% in both cases for the RC images, indicating the neutrality of the normalization model for the reference images. On the contrary, it was a significant improvement in the classification performance when applying the stain normalization to the other centers. Reactive lymphocytes were the most sensitive to stain normalization, with true positive rates (TPR) increasing from 46.3% - 66% for the original images to 81.2% - 97.2% after digital staining. Abnormal lymphocytes TPR ranged from 31.9% - 95.7% with original images to 83% - 100% with digitally stained images. Blast class showed TPR ranges of 90.3% - 94.4% and 94.4% - 100%, for original and stained images, respectively.

Conclusions: The proposed GAN-based normalization staining approach improves the performance of classifiers with multicenter data sets by generating digitally stained images with a quality similar to the original images and adaptability to a reference staining standard. The system requires low computation cost and can help improve the performance of automatic recognition models in clinical settings.

☆ This document is a result of the research project "Computational hematopathology: deep learning solutions for the diagnosis of hematological diseases from peripheral blood cell images (CellsiMaticDeep)". Funded by the Ministry of Science and Innovation of Spain, PID2019-104087RB-I00.

[*] Corresponding author.

E-mail addresses: kevin.barrera@upc.edu (K. Barrera), jose.rodellar@upc.edu (J. Rodellar), santiago.alferez@upc.edu (S. Alférez), amerino@clinic.cat (A. Merino).

https://doi.org/10.1016/j.cmpb.2023.107629

1. Introduction

Peripheral blood (PB) is a tissue that circulates throughout the human body providing oxygen and fighting infections. Accessing this tissue is easy and minimally invasive. PB contains leukocytes, erythrocytes and platelets. Normal leukocytes are divided in neutrophil granulocytes, eosinophils, basophils, lymphocytes and monocytes, being in charge of fighting bacteria, viruses or other foreign substances. An abnormal leukocyte maturation process results in an uncontrolled invasive proliferation of abnormal cells known as cancer. Morphological analysis of cells circulating in PB is the starting point for the diagnosis of more than 80% of hematological [1] diseases. In this analysis, images of cells are obtained that are identified according to their characteristic morphology by clinical pathologists.

The PB morphological analysis process begins by taking a blood sample from the patient and placing a drop on a slide to prepare a blood smear. Blood is spread with another slide to be dispersed on the smear creating an area where cells are observable. In the next step, standardized staining is performed by adding concentrations of chemicals to enhance the contrast of the cell image and highlight the morphological features of the cells. Finally, digital images of the cells are obtained under a microscope with a camera or using digital analyzers, such us the CellaVision DM96 (CellaVision, Lund, Sweden). The clinical pathologist identifies the morphological differences between the different classes of cells circulating in blood, observing parameters such as: cell number, size, nucleus/cytoplasm ratio, nuclear contour, chromatin pattern, cytoplasmic size, granulation and inclusions [1].

May Grünwald-Giemsa (MGG) [2] is one of the most widely used stains for blood smears and bone marrow. Based on the Romanowsky staining procedure, MGG is composed of eosin (an acid dye), methylene blue (a basic dye) and related azures (also basic dyes). These substances are responsible for producing various colors, which allow us to distinguish the following morphological aspects of nucleated blood cells: 1) cell nuclei and chromatin residues (purple); 2) cytoplasm of lymphocytes (blue); 3) cytoplasm of monocytes (gray-blue); 4) granules of basophils (dark blue); 5) granules of eosinophils (reddish to red-brown). Eosine and methylene blue are very sensitivity to pH variations in cellular structures. Each clinical laboratory has its own staining protocol, depending on factors such as the proportion and concentration of the components, the duration of their contact with the smears, and the agents supplied by different manufacturers. Consequently, there is a variability in staining performance between laboratories, which can induce variations in color of cell images, together with other optical effects associated with the illumination of the slide and the quality of the microscope lens. This variability can lead to inconsistencies between the diagnoses made by clinical pathologists [3]. To address this issue, stain normalization has emerged as a promising solution to reduce variability and improve the quality of digital images [4]. Recent research findings [5] have demonstrated that automatic manipulation of color channels can be a valuable tool in improving pathological perception from a clinical standpoint.

The combination of morphological analysis and artificial intelligence (AI) based on medical instruments has become increasingly popular in recent years. This combination makes it possible to recognize qualitative patterns of cell morphology under the supervision of expert clinical pathologists to aid in the diagnosis of hematological and non-hematological diseases. AI algorithms analyze images to perform automatic classification. The variability of staining in images is crucial to the performance of automatic recognition systems [6]. In fact, a system trained for a certain task using images from a specific laboratory may show reduced performance when used for the same task with similar images obtained from another laboratory. This problem is the starting motivation for the approach presented in this paper, which is based on the use of Generative Adversarial Networks (GANs).

Our group has experience in the development of deep learning models to help in the morphological diagnosis of diseases such as leukemia, lymphoma, myelodysplastic syndrome, or malaria parasites [7–10]. These models were trained using image sets annotated by expert clinical pathologists and obtained from fixed and stained smears following a MGG procedure established in the Core Laboratory of the Hospital Clinic of Barcelona (Spain) [11], which is the Reference Center (RC) for this study. Some preliminary tests showed decreased classification performance when using some models to recognize images from other hospitals, which was associated with inter-laboratory variability in color staining among laboratories. In fact, color-based features of blood cell images are among the most relevant to automatic classification systems [12]. Therefore, color variations related to staining protocols become an issue that must be addressed to allow models trained at a specific center to be transferred to other centers.

The first idea was to design a model with a single GAN structure where the input could be color images of other centers different to RC. Conceptually, GANs use two groups of images, source Z and destination (target) Y. A generator G learns to approximate the distribution from Z to Y, resulting in a set X of synthetic images. A discriminator D distinguishes the realism of images X versus Y. In our work, the model should approximate the color distribution of input RGB blood cell images to the color distribution of the RC. That is, obtain synthetic images (X) acquiring the RC staining style while maintaining the same original cell morphology.

Initially we explored the use of VanillaGAN, Deep convolutional generative adversarial networks (DCGAN), Conditional GAN (CGAN) and Wasserstein GAN (WGAN) architectures. However these architectures use random noise at the Z input. The staining of the resulting images X was similar to the RC style but the morphology of the source cells was randomly modified. The use of image translation architectures such as CycleGAN to approximate the staining of the source images to the staining of the RC was also explored. However the result was unsatisfactory, as the resulting synthetic images had a combination of both stains. The trial of pixel-to-pixel translation of the PIX2PIX GAN architecture considerably improved the result, with X maintaining the cell morphology of source Z and a similar staining when using images from the RC. However, the approximation to the RC staining was worse when this idea was used with input images collected in centers other than RC.

To mitigate this variability, we considered changing the color space of the RGB images Z to grayscale prior to performing the pixel-to-pixel translation. This means to translate a single component instead of three. To do this, first a global linear transformation for grayscale was used to obtain Z_{gray} images. The application of the PIX2PIX GAN architecture to approximate the staining distribution of Z_{gray} images to the RC standard improved the result compared to previous tests. But when using Z images with markedly different staining from other centers, the approximation was worse.

Another idea was to focus on the transformation from RGB to grayscale. The global linear transformation is a simple equation with fixed parameters and could be replaced by a more general ”adaptive” model. With this idea, we considered another PIX2PIX GAN to carry out such transformation and this left to the scheme with two sequential GANs called Stain Normalization Model (SNM) shown in Fig. 1.

The contributions of this work are the following:

1. We propose a new two-module system for the normalization of color staining of PB cell images, so that it transforms stained images from different centers to map the color staining of a

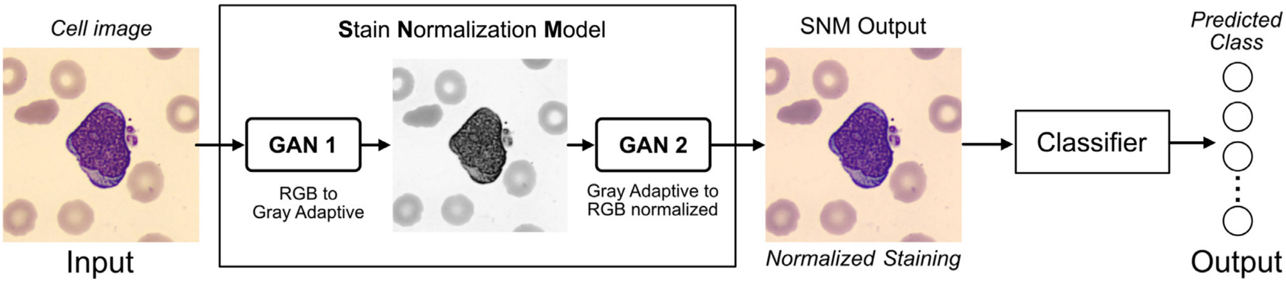

Fig. 1. General structure of the Stain Normalization Model (SNM) in conjunction with a classifier for images of peripheral blood cells. GAN 1 is used to convert original color images to an intermediate gray scale, while GAN 2 learns the standard color staining of the reference center to generate new normalized color images.

reference center while preserving the structural morphological features.

2. The GAN 1 is in charge of mapping each incoming pixel for each color component to an adaptive one-dimensional gray space. Once the incoming images have been decolorized, GAN 2 converts them to RGB normalized images, which are ready for classification.
3. We evaluated the digitally stained images in two ways. The first is a quantitative evaluation of the digitally stained images obtained from the normalization model using GAN metrics. The second uses histograms to have visual insight and interpretations on how the system modifies the color spaces without altering morphology.
4. We also evaluated the effectiveness of the normalization system as a pre-processing tool before the cell images go through a classifier trained with digital images from the RC. For this evaluation, a new CNN classification model was specifically designed for three cell classes of interest: Blasts (BL), Abnormal Lymphocytes (ALC) and Reactive lymphocytes (RL). These are representative of three broad cell types related to leukemia, lymphoma and infection, respectively.

1.1. *Related works*

Generative adversarial networks (GANs) are deep learning models that learn patterns and features from sets of digital images for artificial image generation, segmentation or translation, among other applications. The operation of GANs is based on two networks, known as generator *G* and discriminator *D*, respectively. GAN learning begins when *G* creates artificial images *X* from learning a set of real data *Y* and *D* takes care of differentiating the created images from the real ones. GAN learning ends when the Generator *G* creates artificial images *X* that manage to trick the discriminator *D* by classifying the images *X* as real.

GANs have different architectures depending on their application. The creation of images using GANs consists of converting a random vector Z to a data distribution Y. VanillaGAN [13] uses a basic architecture of discriminator and generator to fit a random data distribution to the data distribution of a real set. DCGAN [14] has a more stable training between generator and discriminator when using convolutions. CGAN [15] uses labels in the discriminator to improve classification performance and WGAN [16] improves training stability using the Wasserstein metric to approximate the random data to real data. Image-to-image translation using GANs consists of approximating a data distribution of a set of images Z to a distribution of a set of images Y. The CycleGAN [17] transforms the domain of an image Z in a different Y using using a couple of generators and discriminators. SRGAN [18] uses a similar concept to improve the quality of images from low to high resolution. StarGAN [19] transforms an image Z domain into multiple Y domains and PIX2PIX [20] uses pixel-to-pixel translation to transform an Z image space into a Y image space by using more robust architectures in the generator and discriminator.

The implementation of GANs to manipulate medical images has increased considerably. The work in [21] presents an overview of GANs in medical applications. In [22] the use of GANs is proposed to modify color combinations in images unrecognizable to people with color vision deficiency. The problem of color staining normalization using GANs has been addressed mainly for histopathological samples. The work in [23] modifies the staining of renal pathology images using the CycleGAN architecture to improve glomerulus detection. In [24] the authors use a CycleGAN architecture to modify the staining of spots in tissues from histological images, evaluating model performance with GAN metrics and in [25] developed a model to learn the color mapping between source and target images for more efficient histological staining computationally. The work of [26] uses a GAN architecture to generate virtual staining on histological images with hematoxylin-eosin concentrations. The work in [27] proposes the use of GANs to transfer the staining of hispathological images of breast cancer and a CNN to segment the smear. Finally, [28] proposes the use of two GANs to carry out a transfer of histological staining.

The use of GANs to the manipulation of peripheral blood cells has been recently considered. In the case of generation of artificial cell images from noise, three recent works are particularly relevant [29–31]. The work in [29] combines autoencoders with StyleGAN to generate four types of leukocytes. In [30] WGAN is used for a similar problem. The paper [31] uses two GAN architectures (WGAN and SRGAN) to generate artificial images of leukocytes and leukemic cells from a random vector.

For the case of stain normalization in blood cells, two recent papers are closely relevant to our work. In [32] authors use the CycleGAN architecture to translate the staining style between two sets of images of leukocytes. This approach produces a morphological modification in the cells, what can affect the image interpretability. In [33] they use a StarGAN-based architecture to digitally stain images of cells obtained with a differential interference contrast (DIC) microscope, which are completely unstained. This is a nice idea in view of robust classification, but DIC images are difficult to interpret by pathologists. In this paper we focus on cell images previously stained with a technique as MGG that is familiar to most of pathologists and extensively used in clinical laboratories. We propose digital color normalization of this broad class of cell images to the color standard of a reference center that is effective for automatic recognition of cells from different centers with reasonable performance.

2. Materials

To develop and assess the effectiveness of the proposed Stain Normalization Model (SNM), we used a dataset of digital images that includes various categories of peripheral blood cells from the following laboratories:

Table 1
Summary of images collected from different centers: RC - Hospital Clinic of Barcelona, Spain; C1 - Hospital Germans Trias i Pujol, Badalona, Spain; C2 - Donostia University Hospital, Donostia, Spain; C3 - Cellavision, Lund, Sweden; C4 - Hospital Josep Trueta, Girona, Spain. Three cell types are considered: abnormal lymphocytes (ALC), blasts (BL) and reactive lymphocytes (RL).

Center	Number of images	Cell Categories			Staining			Cell analyzer	Size of images
		ALC	BL	RL	Blood collection	Protocol	Sample processing		
RC	44,822	17,437	25,062	2323	EDTA as anticoagulant	May Grünwald Giemsa staining	AutomaticSystem	CellaVision DM96	360x363
C1	163	47	80	36					
C2	1506	751	236	519					
C3	547	115	376	56					
C4	925	414	302	209					

RC: Hospital Clinic of Barcelona, Spain
C1: Hospital Germans Trias i Pujol, Badalona, Spain
C2: Donostia University Hospital, Donostia, Spain
C3: Cellavision, Lund, Sweden
C4: Hospital Josep Trueta, Girona, Spain

Blood samples were collected in EDTA as anticoagulant during the daily work of the laboratories. They were automatically stained using the SP10 automatic system (Sysmex, Kobe, Japan) [34] and May Grünwald-Giemsa staining [35]. Digital cell images were obtained from PB samples using the CellaVision DM96 system from each laboratory. The system uses a motorized optical microscope and proprietary software that takes images of the smear and segments them into individual images with a resolution of 360 × 363 × 3 pixels [36]. Clinical pathologists identified and stored the cells of interest for this study based on their morphological characteristics. The diagnoses of the patients were confirmed by integrating all the complementary information such as clinical data, morphology, flow cytometry, cytogenetics and molecular biology [37]. Digital images were labeled as ground truth according to the three cell groups of interest for this investigation: abnormal lymphocytes (ALC), blasts (BL) and reactive lymphocytes (RL). It covers three broad clinical cases: lymphoma, acute leukemia and infection, respectively.

Table 1 presents a summary of images collected from all centers. The total number of images collected from RC, C1, C2, C3, and C4 were 44822, 163, 1506, 547, and 925, respectively. However, not all the RC images were used for training given the need for balancing the dataset. It is worth noting that the images from C1, C2, C3 and C4 were exclusively used for testing the performance of the SNM system, and were not used as part of the training set. The subsequent section details the specific subset of RC images that were employed for both training and testing purposes.

2.1. Reference center images dataset

The RC images were used for training the SNM, as our goal was to approximate the color spaces of other datasets to that of the RC staining, which served as a reference. This approach was taken to address the challenge of color variation among different datasets. The RC dataset consisted of a total of 44,822 images and was divided into two groups following the 80/20 rule [38]: 80% of the dataset where used for the system development and its images were randomly divided into training and validation sets with the same 80/20 rule; the remaining 20% was reserved for the final evaluation (image testing set). The training and validation sets were used to train the two GAN models, and also the classifier used in the experimental evaluation. Additionally, the training of the GAN models was performed using the concept of pixel-to-pixel translation that is presented in detail in Section 3.2.1. This required duplicating the sets by changing the original color of the images to grayscale to obtain paired sets in two different color spaces [39]. The RC images dataset numbers are given in Table 2.

Table 2
Peripheral blood cell images used in this work to develop the Stain Normalization Model (SNM) and the classifier used in this work. Images are grouped by class for each data set.

Cell Type	Number of images			
		Balanced image database		
	Database	Training	Validation	Testing
Abnormal Lymphocytes (ALC)	17,437	2973	530	3470
Blasts (BL)	25,062	2970	513	5344
Reactive Lymphocytes (RL)	2323[a]	2970	527	465
TOTAL	44,822	8913	1570	9279

[a] This group was up-sampled for training with data augmentation techniques to balance all classes.

In general, deep learning requires a database with a large number of samples. In previous works [7,31,40] it was shown that balancing classes is effective in stabilizing training with high precision. We balanced the different classes using classical data augmentation techniques [41] implementing small random transformations: vertical and horizontal twists, zoom, image rotations, and brightness variations [42]. In the end, the training and validation sets were balanced with almost the same number for each cell class, as seen in Table 2. The original test set was used to assess the performance of both SNM and classifier using images not used in training.

3. Methodology

3.1. Overview

It is initially assumed that there is a machine learning or deep learning model for the classification of digital images of PB cells. This model has been trained and tested using cells from smears prepared under a standardized staining protocol at a Reference Center (RC) and images have been obtained using a specific microscope-based analyzer. The objective of this work is to exploit the potential of Generative Adversarial Networks (GAN) to design a new model that provides the classifier with a degree of robustness against variations in staining conditions when images are collected from other hospitals or centers.

Fig. 1 shows the proposed Stain Normalization Model (SNM) as a previous step within a classification scheme. SNM is designed through two sequential steps. In the first step, GAN 1 is responsible for converting the input image pixels from three RGB color components to a grayscale image. This serves to reduce the variations of the three colors resulting from the smear staining, assigning each color a value in a gradation of gray. In the second step, GAN 2 learns the standard staining of the RC image sets and transforms the pixels from gray to three normalized RGB color components. The objective of this transformation is to obtain colored images with an approximation to the color distribution of the RC database.

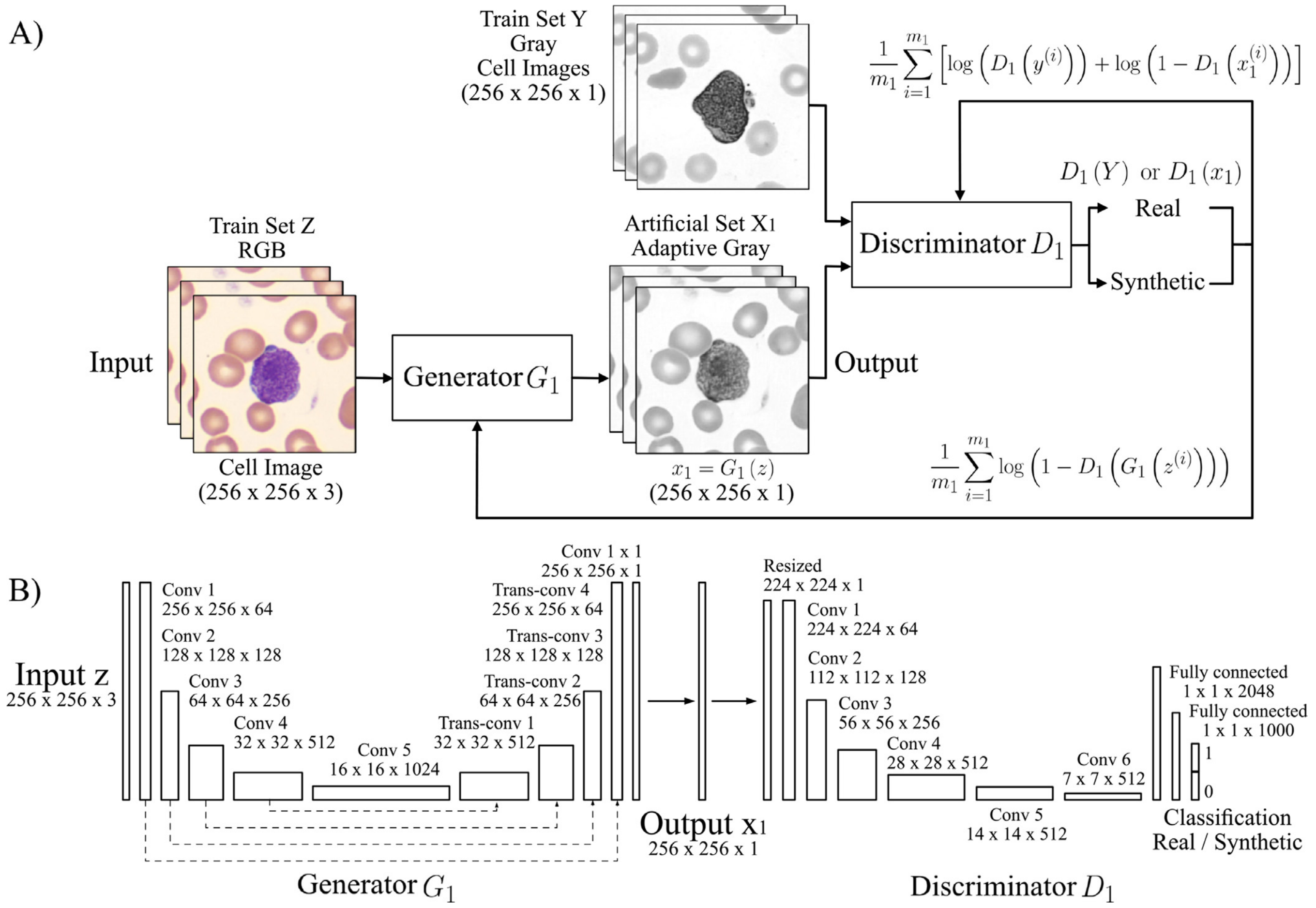

Fig. 2. Architecture of GAN 1. A) The inputs are RGB blood cell images. Generator G_1 creates synthetic images in grayscale. In the training stage, Discriminator D_1 is used to classify the synthetic images as real or fake. D_1 is trained to be the best possible classifier using the train set Y with real gray-scale images as a reference. On the other hand, G_1 is trained until it defeats the discriminator. Then, the system is ready to produce images of cells in an adaptive gray scale. B) Structure of the generator G_1 and discriminator D_1 .

Although the SNM model is general, a prototype case is adopted in this paper to help in the explanations and in carrying out various evaluation experiments. Three cell groups are considered: abnormal lymphocytes (ALC), blasts (BL) and reactive lymphocytes (RL).

The remainder of this section describes the methodological steps to build the trained SNM. Sections 4 and 5 present two experimental evaluations of the quality of normalized images. Section 6 describes the evaluation of the SNM when used in conjunction with a classifier for the three cell groups of interest. This classifier is specifically designed and trained in this paper using RC images only. The SNM evaluation performs the classification of cell images from four different sources.

3.2. Structure of the stain normalization model

3.2.1. GAN 1: From RGB to adaptive grayscale

The PIX2PIX technique is used, where the objective is to translate a source image into an image that plausibly belongs to a certain class. As illustrated in Fig. 2.A, the model is trained using two paired datasets [43]. In this work, one is the source set of RGB images of blood cells (see Table 2) and the other (target) is the duplicated set of the same grayscale images.

As seen in Fig. 2.A, for all source images z, Generator G_1 creates the synthetic grayscale images $x_1 = G_1(z)$, while Discriminator D_1 is a binary classifier that discerns whether the new images are real grayscale images or synthetic images [44]. The generator G_1 is a U-NET convolutional neural network with ResNet. The U-Net (totally convolutional network) architecture was developed for the segmentation of biomedical images [45]. It has U shape due to the three sections that make up the network: contraction, bottleneck and expansion (see Fig. 2.B). In the contraction stretch, the input image z is a matrix of size $m \times n \times 3$, where m is the width, n is the height, and 3 are the RGB color channels. This matrix is resized to $256 \times 256 \times 3$, and goes through five convolutional blocks. Each block applies convolution filters of 3×3 followed by a maximum pooling of 2×2, thus effectively learning complex structures, similar to a CNN. This process is carried out until a feature map matrix of $16 \times 16 \times 1024$ is obtained. This stretch is known as the bottleneck, the lowest layer covering the space between the contraction stretch and the expansion stretch.

In the expansion stretch, the feature matrix goes through four blocks of transposed convolutions to increase the size. In each block the matrix passes through 3×3 filters followed by a 2×2 upsampling layer where the grouping operations are replaced by upsampling operators, thus causing the layers to increase in size.

The stretches of contraction and expansion are interconnected to extrapolate the values in the pixel-to-pixel translation process. From the first convolution, the feature maps are obtained and allocated to their corresponding transposed convolution. In this way, a symmetry is maintained, ensuring that features are learned when compressing and reconstructing the image influenced by the grayscale image set Y. In the last layer of the generator, a 1×1 convolution is applied to each pixel to reduce the depth of the ar-

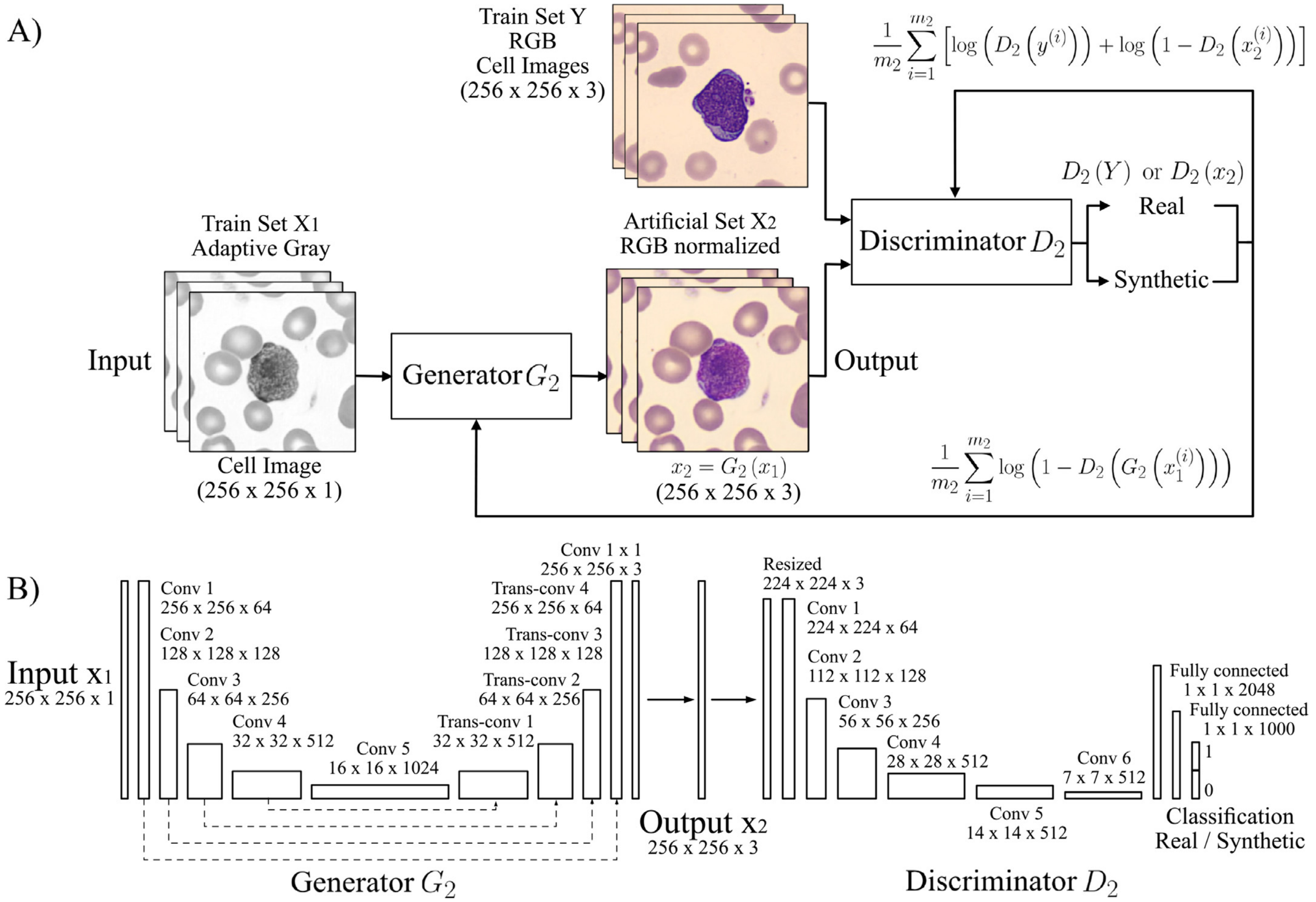

Fig. 3. Architecture of GAN 2. A) The inputs are grayscale blood cell images synthesized by GAN 1. Generator G_2 creates synthetic RGB images. In the training stage, Discriminator D_2 is used to classify the synthetic images as real or fake. D_2 is trained to be the best possible classifier using the train set Z with real RGB images as a reference. On the other hand, G_2 is trained until it beats the discriminator. Then, the system is ready to produce images of cells in RGB color space normalized to our database. B) Structure of the generator G_2 and discriminator D_2.

ray, outputting an image of size 256 × 256 × 1. Through this process, the U-Net architecture learns to assemble a precise output.

The selection of an appropriate architecture is crucial to maintain the fidelity of the input images when using image-to-image translation techniques. The use of a U-Net generator and a PatchGAN discriminator is a traditional approach in the pix2pix architecture [20]. While PatchGAN is a viable technique to identify a specific type of texture in image patches, you may experience problems with more complex shape changes, which may affect the stability and accuracy of each patch, and a loss of global information in the image [46]. In the case of cell images, it is necessary to preserve the morphological information of the entire image. An architecture that considers the whole image and that has proven to be effective in medical image processing is the ResNet34 architecture, uses residual layers to allow lossless information propagation [47]. Consequently, in this work the discriminator D_1 is a binary classifier based on a convolutional neural network with ResNet34 architecture, which labels the real grayscale images and the synthetic ones created by the generator G_1 and decides if the image is real or synthetic. The generator is adjusted based on this decision [48].

The ResNet34 architecture has a normalized input size, so the images x_1 in the generator were resized from 256 × 256 × 1 to 224 × 224 × 1. Consistently, the size of the Y grayscale images were resized from 360 × 363 × 1 to 224 × 224 × 1. All these images go through six convolutional blocks until a feature vector of 1 × 1 × 2048 is obtained. Then two fully connected layers are implemented to reduce the vector. In the last layer a sigmoid activation function gives the binary classification (real or synthetic) with a probabilistic value.

3.2.2. GAN 2: From gray adaptive to RGB normalized scale

GAN 2 has a similar structure to GAN 1 but with an opposite objective. As illustrated in Fig. 3.A, the source images are now those synthetic images x_1 that were generated by GAN 1. The paired data set (Z) is now our RGB image dataset (Z) with the RC staining. Input images x_1 go through Generator G_2 with the same U-Net architecture, and create synthetic RGB images $x_2 = G_2(x_1)$. The D_2 discriminator is a binary classifier that discerns whether the generated colour image is real or not.

The next section presents how both GAN 1 and 2 were trained. After the training was completed, the SNM was ready to operate as shown in Fig. 1. In the final operational setup, only Generators 1 and 2 remain to perform the staining normalization prior to the classification step. Discriminators play their role only in training the GANs.

3.3. Training the stain normalization model

The augmented training set and the validation set in Table 2 were used for the learning of the SNM. Since we used the PIX2PIX image-to-image principle, a grayscale transformation was applied on the original RGB images to have two paired training

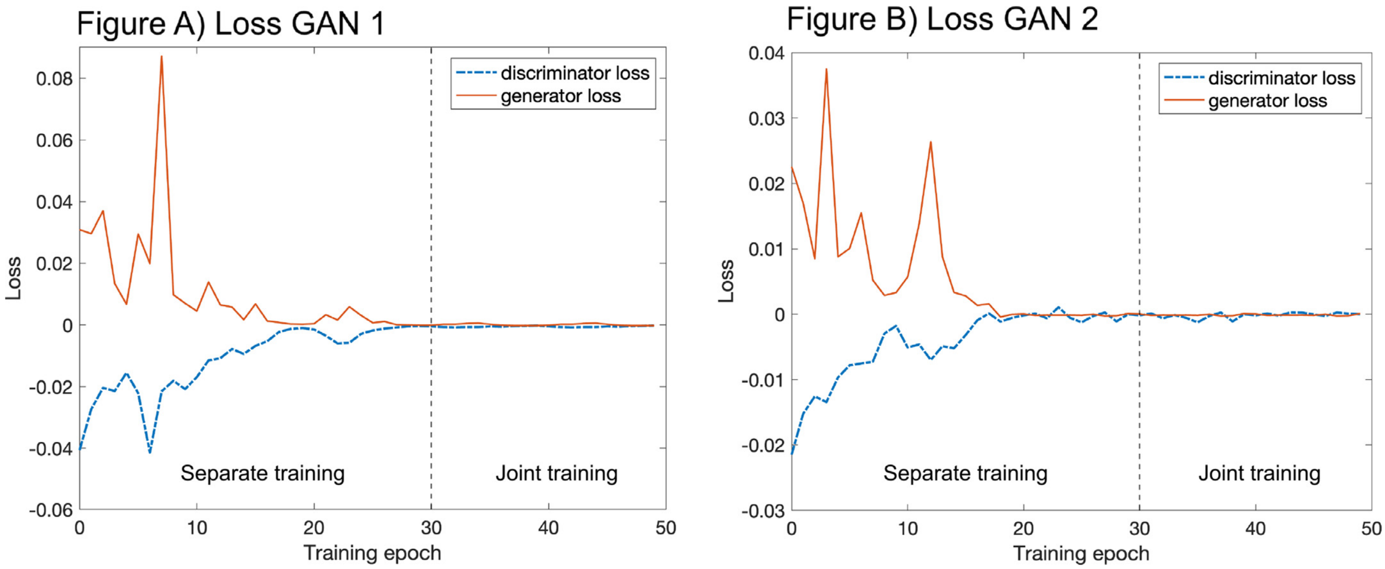

Fig. 4. Loss progression during the system training. Each network was trained for 50 epochs. Figure A represents the losses of GAN 1, responsible for modifying the RGB input to adaptive gray. Figure B represents the losses of GAN 2 that modifies the adaptive gray space to normalized RGB.

sets, which were used as illustrated in Fig. 2.A and 3.A. A server with a 12 GB Nvidia Titan XP GPU was used in this work. We discuss computational resources in detail in Section 7.

In this section, the training of GAN 1 is described. GAN 2 was similarly trained. Traditional GAN learning methods are based on simultaneous training of the generator and the discriminator. However, this procedure has limits in the GPU memory, longer processing time, instability in the results and low training control. Alternatively, we opted to use the NoGAN training method [49]. It consists of training the generator and the discriminator separately in an initial stage, having a greater control in the training, reducing the GPU memory and optimizing the learning time. After this stage, additional joint training took place. The training is counted in epochs. An epoch is a hyperparameter [50] that indicates the period in which all the images of the dataset pass through the network for model training. We iteratively processed batches of 40 images until completing the total number of images in each epoch to improve learning efficiency. In this process, it is necessary to optimize the generation and discrimination error, so we use traditional GAN loss functions.

First the generator is trained separately, using the RGB and gray-scale training sets, Z and Y respectively. At each learning epoch, the generator G_1 creates synthetic images $G_1(z^{(i)})$ where $z^{(i)}$ is the $i-th$ RGB input image. In this stage, all the parameters of the discriminator D_1 are frozen and set with initial default values. The discriminator D_1 was designed to be a classifier with an output vector of two elements belonging to the labels real and synthetic.

Initially, the probability $D_1(G_1(z^{(i)}))$ is closer to 1 in the synthetic label. The generator was trained to fool the discriminator, so the probability $D_1(G_1(z^{(i)}))$ must decrease in the synthetic label. Consequently, this objective was stated as the minimization of the following loss function:

$$L_{G_1} = \frac{1}{m_1}\sum_{i=1}^{m_1} \log\left(1 - D_1\left(G_1\left(z^{(i)}\right)\right)\right) \tag{1}$$

where m_1 is the number of images generated by G_1.

For the implementation of the system training code, we used FastAI [51]. It is a deep learning library for the development and research of artificial intelligence algorithms, which uses the Pytorch numerical calculation package of the Python programming language. From FastAI we used the *GANLearner* library, which has the *show_img* tool that was activated to view the images $G_1(z)$ and follow the generator learning progression. In each epoch of training the generator with the NoGAN technique, we used the *GANLearner* library to obtain the produced generator loss defined in Eq. (1), whose progression is displayed in Fig. 4 (A).

After conducting several tests, using the *show_img* function to evaluate the quality of synthetic images generated by the GAN and analysing the behaviour of the generator loss curve, we determined that 30 training epochs using the NoGAN method were optimal for our purposes. We selected this number of epochs based on the observation that the generator loss curve was closest to 0 before destabilizing. Additionally, we performed a visual evaluation of the generated images and determined that the generator's performance was satisfactory. At the end of the generator training, a set X_1 of synthetic gray scale images was produced.

Next we used the set X_1 together with the set Y of real grayscale images for separate training of the discriminator as a binary classifier. In this case, we considered two probabilities $D_1(x_1^{(i)})$ and $D_1(y^{(i)})$. When inputting an image $y^{(i)}$, the probability $D_1(x_1^{(i)})$ must be close to 0 and the probability $D_1(y^{(i)})$ must be close to 1. When inputting an image $x_1^{(i)}$ the probability $D_1(x_1^{(i)})$ must be close to 1 and the probability $D_1(y^{(i)})$ must be close to 0. This means being accurate in recognizing images in set Y as real, while recognizing images in set X_1 as synthetic. Consequently, this objective was established by maximizing the following loss function:

$$L_{D_1} = \frac{1}{m_1}\sum_{i=1}^{m_1} \left[\log\left(D_1\left(y^{(i)}\right)\right) + \log\left(1 - D_1\left(x_1^{(i)}\right)\right)\right] \tag{2}$$

where m_1 is the number of images in sets X_1 and Y.

Fig. 4 (A) shows the D_1 loss progression defined in equation (2). After 30 epochs of training, D_1 reached an accuracy of 96.2% in the binary classification, we considered to be a satisfactory level of performance.

After training the generator and discriminator separately, an additional joint training stage was carried out. In each joint training epoch, the weights of the generator G_1 are updated and a set of images x_1 is created. This set is classified by the discriminator D_1 and its weights are updated. The loss functions of the generator (1) and the discriminator (2) are optimized independently in the same training epoch. This process was repeated during several epochs of joint learning. At epoch 20 of joint training, the $x_1 = G_1(z)$ images created by the generator beat the discriminator. The evaluation visual of these images indicated that the morphology of the generated cells was visibly similar to that of the original set Y, thus ending the GAN 1 training. Fig. 4 (A) shows that the joint training (20 epochs) maintained the stabilized low values of both the generator and the discriminator losses. The complete training of GAN 1 was 50 epochs.

The entire process described to train GAN 1 was performed to train GAN 2. As highlighted in Fig. 3, the only difference is that we used the grayscale set X_1 (generated by GAN 1) and the RGB set as paired training sets. To summarize the training result, Fig. 4 (B) shows the progression of the loss functions for the Generator G_2 and the Discriminator D_2, respectively. We may observe that GAN 2 also stabilized its losses after the 50 training epochs. Fig. 4 shows that both GAN models did not present overfitting, so the weights obtained after the described training determined the final SNM model.

4. Quantitative evaluations of the SNM

We considered examining the staining effectiveness of the SNM using a quantitative evaluation of the synthesized images. Peripheral blood cell original images of RC, C1, C2, C3 and C4. These images were not used for any training. For RC, the 9279 images of the testing set detailed in Table 2 were used. In addition, 163 images from C1, 1506 from C2, 547 from C3 and 925 from C4 detailed in Table 1. The images had visible staining differences between centers.

We started by organizing the image sets for the evaluation. The set of images with the original color of each center was called Y. We duplicated the set Y of each center to Y_{gray} by changing the color space to grayscale using a global linear transformation. The set Y passed through GAN 1, obtaining the set X_1 in adaptive grayscale. Sequentially, the set X_1 went through GAN 2, obtaining the set X_2 in normalized RGB color.

Two tests were carried out:

A: The normalized RGB X_2 images synthesized by the SNM are compared with the original RGB images Y.

B: The adaptive gray X_1 images synthesized by GAN1 are compared with the original paired gray images Y_{gray}. This means a comparison between the adaptive gray transformation by GAN 1 and the global linear transformation.

The most regularly used metrics in GAN models were calculated: Frechet Inception Distance (FID) [52], Inception Score (IS) [53], and Learned Perceptual Image Patch Similarity (LPIPS) [54]. The Table 3 presents the result of the metrics of each center.

The FID metric scores the distance between feature vectors extracted from two sets X and Y. The feature vector is extracted using an InceptionV3 model pretrained with its original weights and the calculation is performed using the 2048 feature vector of its last pooling layer. Conceptually, the smaller the distance from set X to set Y, the color of X will tend to be similar to that of Y.

The FID score in test A between X_2 and Y is evaluated taking two factors into consideration. First, we consider that the FID score in test A_{RC} should ideally tend to zero since the objective of GAN 2 is to replicate the color of the RC. Second, the FID score in test A for the other centers must be clearly greater than the score obtained for the RC case to confirm the effective color normalization by the SNM. The FID score of the test A_{RC} in Table 3 is 3.532. This is a low value that indicates a small bias in the RGB color distribution between the original RC images and those generated by the SNM. On the contrary, the FID score in test A in the rest of the centers is considerably higher, this being the objective of the SNM. On the other hand, the FID score in test B between X_1 and Y_{gray} also shows a small bias (4.418) for the RC between the classic linear gray transformation and the adaptive transformation by GAN1. On the contrary, the score shows significant higher values for the rest of the centers, which is consistent with the objective of GAN1.

Table 3
Quantitative evaluation metrics for the Stain Normalization Model. A) Comparison between original RGB and synthetic normalized RGB cell images. B) Comparison between adaptive gray transformation and classic gray transformation.

Test	Metrics		
	FID	IS	LPIPS
A_{RC}	3.532	10.552 ± 0.138	0.039
A_{C1}	29.703	10.416 ± 0.076	0.063
A_{C2}	10.199	10.192 ± 0.156	0.043
A_{C3}	9.301	10.585 ± 0.174	0.048
A_{C4}	26.493	10.471 ± 0.133	0.049
B_{RC}	4.418	10.435 ± 0.288	0.068
B_{C1}	47.864	10.261 ± 0.153	0.065
B_{C2}	14.287	10.274 ± 0.082	0.047
B_{C3}	14.566	10.491 ± 0.256	0.061
B_{C4}	31.578	10.341 ± 0.195	0.058

The IS metric is used to score the quality of a set of artificially generated images through the classification of the set with an InceptionV3 model pretrained with its original weights. In this study, the score was calculated for image sets X_2 for test A and X_1 for test B, respectively, for each center. The IS score in Table 3 is similar in the two tests for all cases. This means that GAN 1 and GAN 2 do not change the overall quality of the input images.

The LPIPS metric is used to assess the visual perception of image sets, similar to human sensation. Random patches of size 14 × 14 pixels are created between pairs of corresponding images X and Y. The metric measures the distance between patches. In test A, the score measured the patch distance between image sets X_2 and Y, while in test B the measure was the distance between X_1 and Y_{gray}. The result of this metric is a considerably low value for each center. This indicates a similar visual perception between the original RC images and the corresponding synthetic images.

5. Histogram evaluation and interpretation

In this section histograms are used to get an idea of how the developed Stain Normalization Model modifies the color spaces of the images without affecting cell morphology. The histograms of the red, green and blue components of each image were obtained before and after applying SNM. For each center, the average of the individual histograms was calculated as a single representative histogram of the entire set.

An example of an individual cell from each center is presented in Fig. 5. The image in (a) represents the source image, and the image in (b) shows the image after applying SNM. It is worth noting that morphological changes are not observable. Additionally, the average histograms of the image sets from each center before and after normalization are shown in Fig. 5 (c) and (d), respectively.

When analyzing each color channel separately, we observed that the red channel in all centers (see Fig. 5 (e)) had the highest intensity (255) in most of the pixels. However, after applying SNM, the maximum frequency of the red channel decreased in all centers, as observed in the average histograms. This reduction is visually noticeable in the examples presented in Fig. 5 (a) and (b), in the red blood cells of the background.

In the case of the green channel of the source images (see Fig. 5 (f)), the maximum frequency peak had some variability between centers. However, after applying SNM, the peak was adjusted to a frequency of 0.8, and in the case of C2, the two peaks merged into one, similar to the other centers.

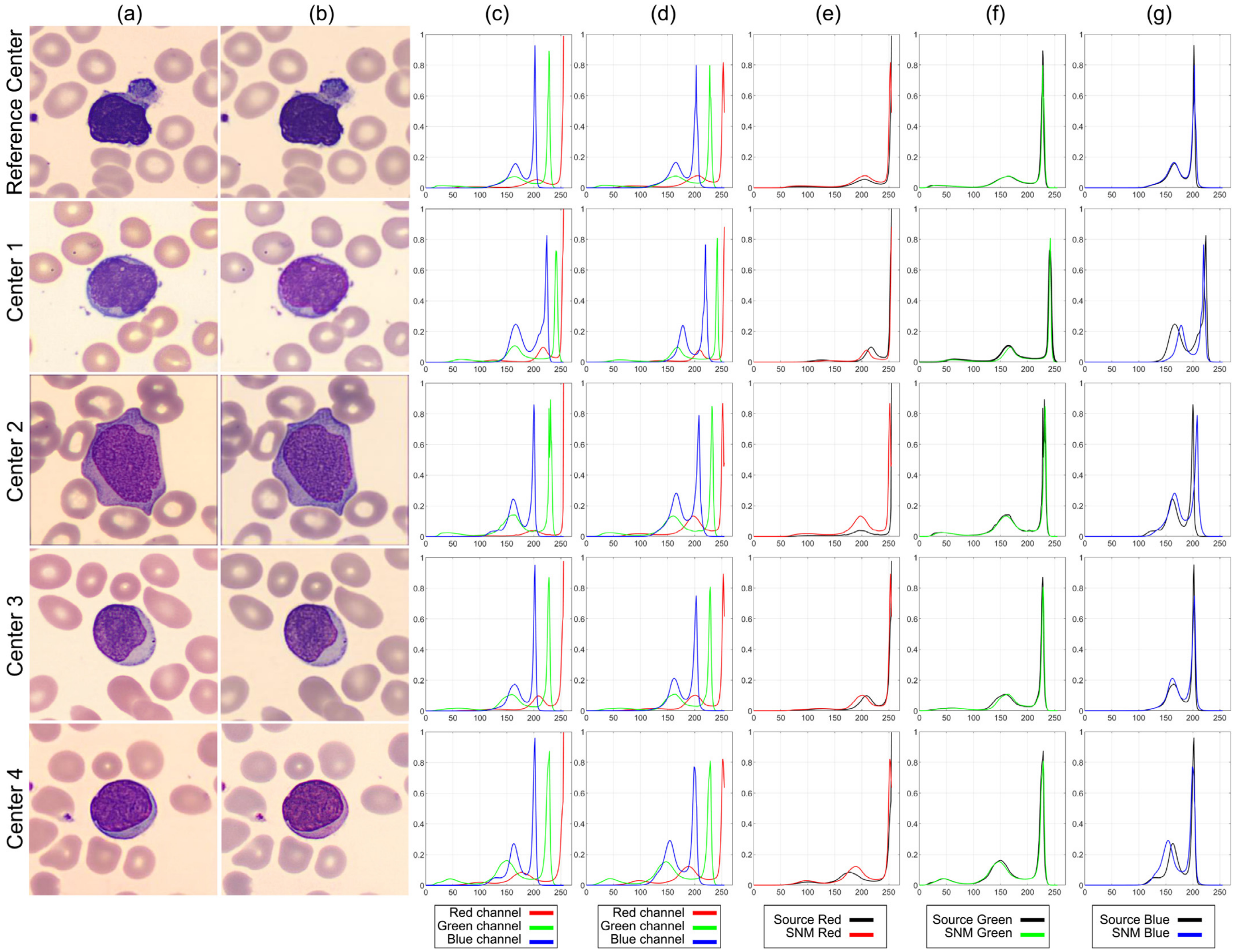

Fig. 5. Image standardization examples with SNM. (a) Original images. (b) Images with standardized staining applying SNM. (c) Average histograms for the RGB components of the image set from each center before normalizing the staining; pixel frequency versus color intensity. (d) The same average histograms after applying SNM. (e) Comparison of the red channel histograms between the original images and images with SNM. (f) Comparison of the green channel histograms between the original images and images with SNM. (g) Comparison of the blue channel histograms between the original images and images with SNM. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Regarding the blue channel of the source images (see Fig. 5 (g)) of the Reference Center (RC), two characteristic frequency peaks were observed. However, in the case of the other centers, SNM modified the frequency distribution of the blue color, achieving a behavior similar to that of the RC. This change is noticeable visually in the cytoplasm of the cells in Fig. 5 (a) and (b).

The color composition in the images can be described as a probability distribution, which is in fact the cumulative color histogram (CCH) defined as follows:

$$CCH_c(x) = \sum_{j \leq x} h_{cj} \tag{3}$$

where c denotes a specific color (for example, red, green or blue), and h_{cj} is the pixel frequency of the image with color intensity j, which is the histogram value for the color c.

It has been argued that CCH is a more robust representation of the color distribution than the histogram for quantitatively determining the similarity between color images [55]. Also, typical L_1 or L_2 metrics give reasonable similarity values simply by calculating the distances between pairs of CCH. With this objective, we calculated the cumulative histograms from the average histograms shown in Fig. 5 (c) and (d). They are shown in Fig. 6 for source images and SNM processed images, respectively, separated by centers and color components.

First, some details can be highlighted from a visual inspection. Analyzing the red channel in Fig. 6 (a) for the source CCHs, approximately 50% of the pixels have an intensity between 200 and 250, while in the SNM CCHs the intensity range decreases between 220 and 230. This is because SNM adjusts the value of the red component to the right and left with respect to the reference staining, and the CCH curves in Fig. 6 (a) join closer to the RC curve. The source CCH and the SNM CCH of the green component have little variation Fig. 6 (b). The blue channel of the C1 CCH source, unlike the rest of the centers, is notably different. Its maximum percentage of pixels is between the range of 225 and 255. When applying SNM, the intensity is adjusted in a range between 200 and 255, moving the blue CCH curve to the left bringing it closer to those of the other centers.

To have a quantitative measure of the previous observations, the CCHs of the RC for source images were used as a reference. For each center and color channel, we calculated the root mean square error (RMSE) between the CCHs and the reference adopted

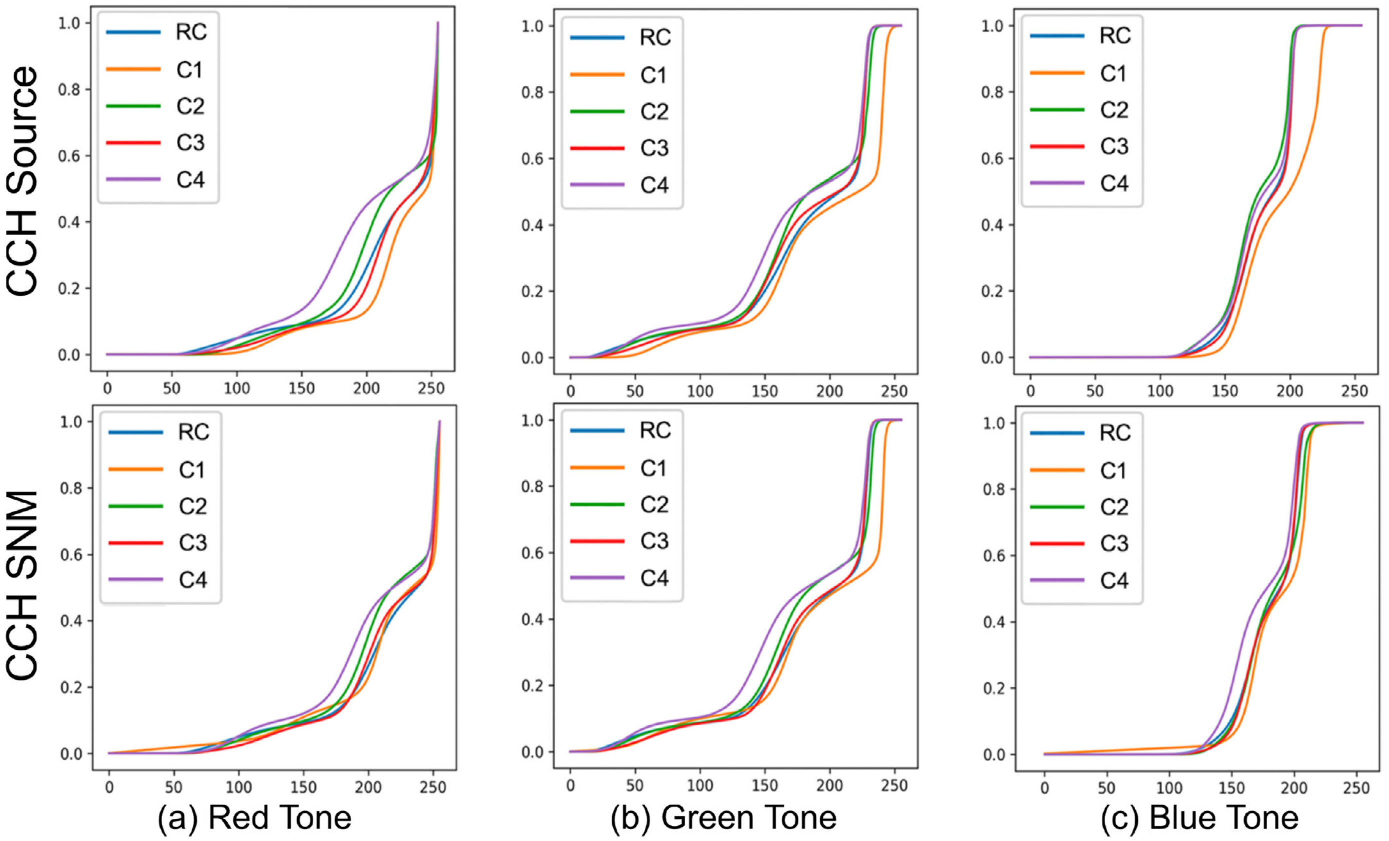

Fig. 6. Cumulative color histogram (CCH) of average histograms of the source staining and SNM staining. CCH for each color channel: (a) red, (b) green and (c) blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

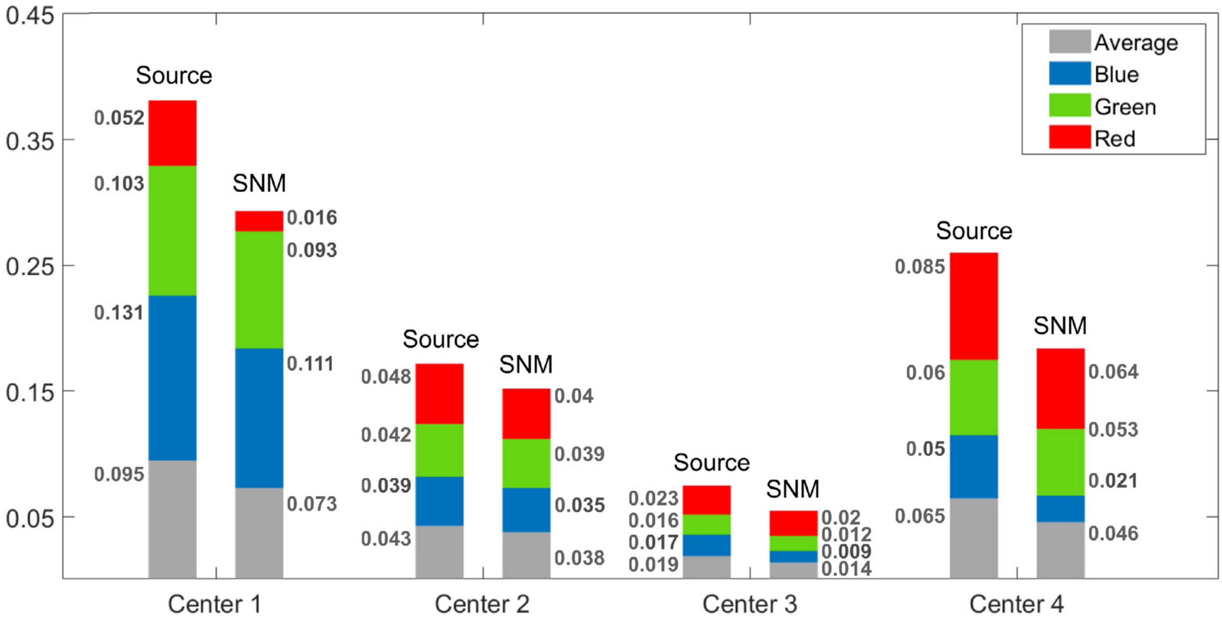

Fig. 7. Root mean square error (RMSE) of the cumulative color histograms (CCH) for each center, before and after using SNM, with respect to the CCH of the source color images of the Reference Center.

before and after using SNM. All these values are shown in Fig. 7. In addition, Table 4 gives the differences between the RMSE before and after applying SNM in percentage relative to the initial source value.

We observe that, in all cases, the RMSE after SNM is lower than for the source images. The greatest relative changes are observed for: (1) the red color for C1, from 0.052 to 0.016, which means a reduction of 69.2% with respect the original value; (2) the blue color for C4, from 0.050 to 0.021, which is a reduction of 58%; and (3) the blue color for C3, from 0.017 to 0.009, which means a reduction of 47.06%.

The bottom row in Table 4 shows that the lowest average reduction per center in RMSE is for Center C2 (11.63%), while it is between 23.16% and 29.23% for the other centers. In terms of col-

Table 4
Relative differences (in %) of the root mean square errors (RMSE) of the cumulative color histograms (CCH) before and after using SNM. The RMSE are calculated between the cumulative color histograms (CCH) of the different centers and the Reference Center.

	C1	C2	C3	C4	Average value per color
Red	69.23	16.67	13.04	24.71	30.91
Green	9.71	7.14	25.00	11.67	13.38
Blue	15.27	10.26	47.06	58.00	32.65
Average value per center	23.16	11.63	26.32	29.23	

ors, the last column of the Table 4 shows that the lowest average reduction per color is for green (13.38%). Corroborating that the red (30.91%) and blue (32.65%) components are more influenced by SNM than the green one.

6. Evaluation of the SNM in a multicenter classification test

The objective of this section is to evaluate the effectiveness of the SNM as a pre-processing tool before the cell images are classified within the scheme described in Fig. 1. First, Section 6.1 presents a CNN model designed specifically in this study for the classification of blood cell images into the following classes: abnormal lymphocytes (ALC), blasts (BL) and reactive lymphocytes (RL). The model was trained on the RC color distribution using the data set shown Table 2. Section 6.2 describes the assessment of this classifier with the test set of RC images (Table 2) before and after using color normalization by SNM. Finally, the most relevant result is the evaluation of the classifier when using images from centers other than the RC (Section 6.3).

6.1. Structure and training of the classifier

Convolutional neural networks have architectures generally separated into two sequential parts: 1) a series of blocks trained to obtain complex characteristics from images; and 2) a block with a fully connected multilayer neural network that is trained to perform the classification using the learned characteristics. We investigated several architectures (Inception [56], PNasNet5Large [57], VGG16 and VGG19 [58]), and SENet154 [59], which were used with satisfactory results in previous works for the recognition of cells circulating in blood. [7–10,40].

Several iterative training and testing experiments were carried out with the database shown in Table 2 and determined the accuracy for the classification of the ALC, BL and RL classes. The resulting testing accuracy for the five CNN models is shown in Table 5. SENet154 was selected for the final assessment based on its highest performance rating.

In the SENet154 architecture, the images of data set where resized to a normalized size of 224 × 224. It is advisable to normalize the data to avoid problems of numerical instability.

The process to train the network is iterative and the time to do it is divided into epochs. An epoch is usually divided into multiple iterations, so that the network processes a batch of training samples at each iteration until the entire training set is used. In this work, we randomly selected mini-batches of 20 samples without repositioning in each iteration. At each epoch, a loss function is used to determine the error, which is a measure of the discrepancy between the prediction made by the network and the true label assigned by clinical pathologists. In this work, the following categorical cross entropy loss function was used:

$$L = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{3} t_j^{(i)} \log\left(\hat{t}_j^{(i)}\right) \tag{4}$$

where m is the number of images in the training set, $t_j^{(i)}$ is the true label of the class and $\hat{t}_j^{(i)}$ is the probability of the predicted class. This means that $t_j^{(i)} = 1$ if the image sample i belongs to the class j, and $t_j^{(i)} = 0$ otherwise.

The objective of the training is to update the network weights to gradually reduce the loss function towards its minimum according to the gradient descent principle. Using the backpropagation approach, the gradients with respect the weights must be determined backward through the network. To optimize the learning process, Adaptive Moment Estimator Optimizer (Adam) [60] was used to estimate the gradients along with the cyclical learning rate policy [50].

After completing each learning epoch, the loss was calculated for the entire training set. In addition, the images from the validation set were classified by the updated model to calculate the loss until the classification accuracy was acceptable. Fig. 8 shows the progress of the training and validation loss and the accuracy of the classifier. Based on the results shown, it was concluded concluded that 20 training epochs were sufficient because at this point the loss was minimal and the validation accuracy reached its maximum value. Therefore, the training/validation stage was completed.

6.2. Classifier evaluation using blood cell images from the reference center

The performance of the classifier is evaluated in two experiments. The first experiment evaluated the performance of the classifier without using the SNM. This means that all 9279 peripheral blood cell images in the testing set detailed in Table 2 were classified without modifying their original RC staining. Table 6 shows the confusion matrix obtained when comparing the predicted classes with the true classes confirmed by clinical diagnosis. Each column represents the percentage of prediction by class and each row represents the actual instances.

From this multi-class confusion matrix, we calculated the sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), precision or positive predictive value (PPV), and $F1$ score for

Table 5
Comparison of convolutional neural network architectures for the classification of ALC, BL and RL.

Images Dataset	Architecture	Training epochs	Testing accuracy
Training	Inception V3	30	92.3%
8913	PNasNet5Large	28	90.9%
Validation	VGG16	25	93.4%
1570	VGG19	25	93.8%
Testing	SENet154	20	94.6%
9279			

Table 6
Confusion matrix of the classification results (in %) for the images of the testing set from the Reference Center. The diagonal shows the true positive values (TP) for each cell class. The balanced accuracy is 95.7%.

RC Original Color		Predicted class		
		Abnormal lymphocytes	Blasts	Reactive lymphocytes
True	Abnormal lymphocytes	**94.4**	3.6	2.0
Class	Blasts	1.3	**97.9**	0.8
	Reactive lymphocytes	3.8	1.4	**94.8**

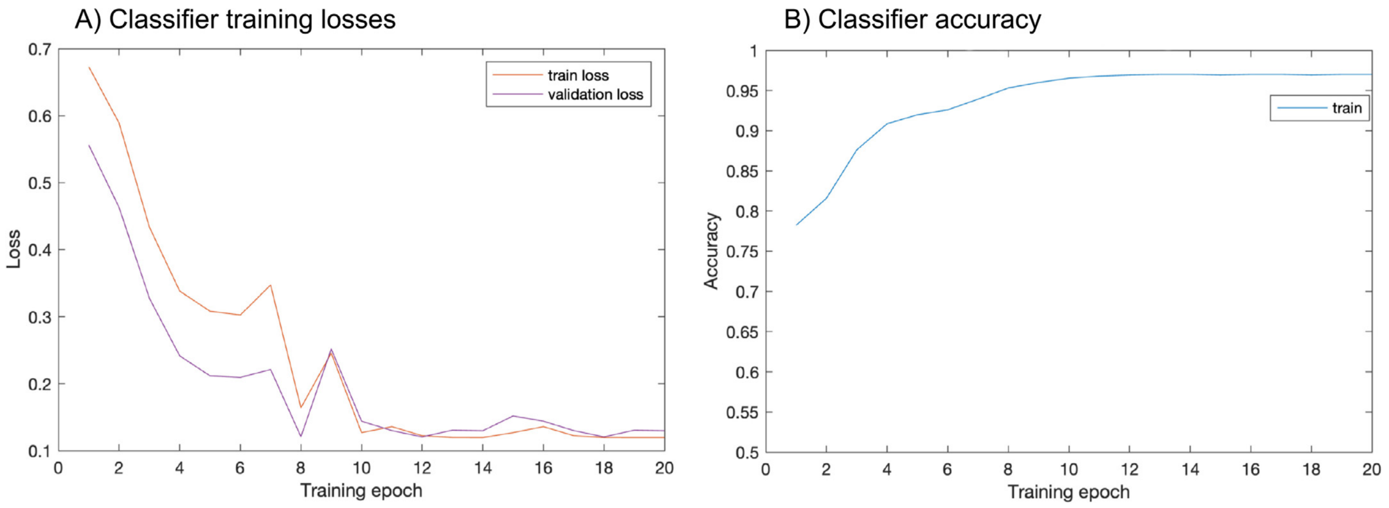

Fig. 8. Classifier training for 20 epochs. A: Progression of training and validation losses. B: Progression of the accuracy for the classification of the images in the validation set.

Table 7
Sensitivity, specificity, precision and F1 score values of the classification results of the images of the testing set from the Reference Center.

	Abnormal lymphocytes	Blast	Reactive lymphocytes	Average values
Sensitivity	0.944	0.979	0.948	0.957
Specificity	0.975	0.975	0.986	0.979
Precision	0.949	0.951	0.971	0.957
F1 score	0.946	0.965	0.959	0.956

Table 8
Confusion matrix of the classification results (in %) for the images of the testing set from the Reference Center after using SNM. The balanced accuracy is 96%.

RC SNM Color		Predicted class		
		Abnormal lymphocytes	Blasts	Reactive lymphocytes
True Class	Abnormal lymphocytes	**95.9**	2.9	1.2
	Blasts	1.8	**97.9**	0.3
	Reactive lymphocytes	3.1	2.7	**94.2**

each cell class i as follows:

$$(TPR)_i = \frac{(TP)_i}{(TP)_i + (FN)_i} \qquad (TNR)_i = \frac{(TN)_i}{(TN)_i + (FP)_i}$$
$$(PPV)_i = \frac{(TN)_i}{(TN)_i + (FP)_i} \qquad F1_i = 2 \cdot \frac{(PPV)_i \cdot (TPR)_i}{(PPV)_i + (TPR)_i} \tag{5}$$

In these expressions, $(TP)_i$ (true positive) is the value in the main diagonal of the confusion matrix. $(FN)_i$ (false negative) is the sum of the row values of class i excluding the diagonal. $(FP)_i$ (false positive) is the sum of the column values of class i class excluding the diagonal. $(TN)_i$ (true negative) is the sum of the entire matrix excluding $(TP)_i$, $(FN)_i$ and $(FP)_i$. In this way we obtain the performance metrics of each class compared to the other two classes. All these values are displayed in Table 7.

Finally, the balanced accuracy is the percentage of images correctly classified, which is the average of TPR for the three classes. Its value was 95.7%. All of the above metrics indicate that the trained model performed satisfactorily for classifying new images that were obtained with the same staining protocol as the images used for training.

The second experiment aimed to evaluate the effect of SNM on the performance of the classifier with the same images. Therefore, the same 9279 cell images from the test set were run through the SNM and then passed through the classifier. Table 8 is the classification confusion matrix of the normalized images with SNM.

It is interesting to compare the results obtained in Table 6 and those obtained with SNM normalization in Table 8. The performance in abnormal lymphocytes slightly improves with SNM, the TPR values increasing from 94.4% to 95.9%. In the case of blasts, there is no variation in TPR, the FN are distributed mainly to the abnormal lymphocyte class. In reactive lymphocytes, a slight TPR decrease is observed (94.8% to 94.2%) when using SNM.

6.3. Classifier evaluation using blood cell images from different centers

In this section, we evaluate the performance of the classifier with images from different centers. Sets of 163 images from C1, 1506 from C2, 547 from C3 and 925 from C4 passed through the classifier in two experiments.

In a first experiment, we passed all the images with their original staining through the classifier and obtained the confusion matrix for each center. In a second experiment, the images went first through the SNM, normalizing their staining according to the color of the RC. Then, the normalized images went through the classifier and we obtained a new classification confusion matrix for each center. All the confusion matrices for both experiments are shown in Table 9.

A first look at the confusion matrices for the original images in Table 9 reveals a significant reduction in the true positive rates (TPR) on the main diagonal for some of cell classes. The lowest TPR is 31.9% for the abnormal lymphocytes from C1. For reactive lymphocytes, these values are 52.8%, 46.3%, 66% and 53.6% for C1, C2, C3 and C4, respectively. These low TPR values are in sharp contrast to the high values shown in Table 6 when using exclusively the images collected in RC.

On the other hand, the confusion matrices on the right side of Table 9 show that the classification of all cell image classes from all centers improves considerably after normalizing the original images using the SNM. For example, the TPR of reactive lymphocytes raised to values of 97.2%, 96.7%, 85.7% and 81.2% for the C1, C2, C3 and C4, respectively. In addition, the TPR of the abnormal lymphocyte class for the C1 increased drastically from 31.9% to 97.9% and the TPR of the RL increased from 52.8% to 97.2% for the C1, and from 46.3% to 96.7% for the C2.

Table 9
Confusion matrices of the classification results (in%). The general classification accuracies before using SNM of Center 1, Center 2, Center 3 and Center 4 were 61.2%, 73.9%, 84.3% and 82.5%, respectively. The general classification accuracies after using SNM of Center 1, Center 2, Center 3 and Center 4 were 98.4%, 92%, 93.4% and 92.6%, respectively.

Center		Original color			SNM color		
C1		Predicted class					
		Abnormal	Blasts	Reactive	Abnormal	Blasts	Reactive
True class	Abnormal lymphocytes	**31.9**	68.1	0	**97.9**	2.1	0
	Blasts	1.2	**98.8**	0	0	**100**	0
	Reactive lymphocytes	8.3	38.9	**52.8**	0	2.8	**97.2**
Global accuracy			61.2%			98.4%	
C2		Predicted class					
		Abnormal	Blasts	Reactive	Abnormal	Blasts	Reactive
True class	Abnormal lymphocytes	**85.2**	14.7	0.1	**83**	13.1	3.9
	Blasts	9.7	**90.3**	0	2.5	**96.2**	1.3
	Reactive lymphocytes	17.3	36.3	**46.3**	1.9	1.3	**96.7**
Global accuracy			73.9%			92%	
C3		Predicted class					
		Abnormal	Blasts	Reactive	Abnormal	Blasts	Reactive
True class	Abnormal lymphocytes	**95.7**	1.7	2.6	**100**	0	0
	Blasts	8.8	**91.2**	0	5.3	**94.4**	0.3
	Reactive lymphocytes	7.5	26.4	**66**	8.9	5.4	**85.7**
Global accuracy			84.3%			93.4%	
C4		Predicted class					
		Abnormal	Blasts	Reactive	Abnormal	Blasts	Reactive
True class	Abnormal lymphocytes	**94**	6	0	**97.3**	2.2	0.5
	Blasts	0	**100**	0	0.7	**99.3**	0
	Reactive lymphocytes	40.2	6.2	**53.6**	17.8	1	**81.2**
Global accuracy			82.5%			92.6%	

7. Discussion

Cell staining is an important step in the preparation of a PB smear for the diagnosis of hematological and non-hematological diseases. May-Grünwald Giemsa (MGG) is a physical cell staining technique widely used in clinical laboratories. It highlights morphological differences in blood cells improving microscopic observation. The current trend in modern clinical laboratories is to use automated systems to improve efficiency in smear preparation, obtain digital images of blood cells, and automatically classify cell classes for diagnosis purposes. CNN-based classifiers are generally trained to achieve performance metrics within some range. Experience shows that this performance is sensitive to variations in the staining process, so new tools are needed to artificially normalize the color spaces without altering the distinctive morphological features of cells.

Within this context, our work uses generative adversarial networks (GANs) to develop a Stain Normalization Model (SNM) to adjust the color spaces of digital images to the standardized color of a reference center (RC). The model operates in two sequential steps. The first removes color from the input image, and the second model digitally stains the image using the RC colors as a target. The resulting images are realistic, preserve cell morphological features and mitigate the performance degradation of CNN-based classifiers against images of cells with variable staining.

Although there are a number of relevant works devoted to color normalization in tissue slides, few articles have been presented in the literature on automatic staining of PB cells using deep learning algorithms. As far as the authors know, the closest to our work are [32] and [33].

The work in [32] uses a CycleGAN to translate the staining style between two sets of images made up of normal leukocytes (neutrophils, eosinophils, basophils, monocytes, lymphocytes), myelocytes, and metamyelocytes. The model is trained on a sample database of 20 patients. The authors mention the overfitting in the loss function that produces a morphological modification in the cells. In our opinion, the preservation of morphology is very important in the staining process, since clinical pathologists and the automatic recognition models rely on the identification relevant features associated to geometry, color and texture. Modifying cell morphology can directly affect the interpretability of the images with a misdiagnosis.

The work in [33] presents a combined digital staining and classification approach that uses unstained cell images obtained with a differential interference contrast (DIC) microscope. Using completely unstained images as input is attractive, as this could imply a classification ideally robust against staining variability. However, extraction of relevant features from DIC images is difficult and auxiliary tasks of segmentation and pairwise reconstruction are needed. On the other hand, DIC images are difficult to recognize through morphological visual inspection. In fact, clinical pathologists are used to working with images of stained cells, with MGG being the most widely used staining method. In this regard, the approach in [33] includes a digital staining method aimed at obtaining colored RGB images with MGG-like appearance and preserved morphology that can be recognizable by expert clinical pathologists.

In our work, the inputs are RGB cell images acquired from patients from different centers using the CellaVision DM96 and stained with their own MGG protocol (with different staining times or pH), which influences the color of the cell images. The final destination of these images is to go through an automatic recognition model previously trained and validated with this type of images but obtained in the RC. The contribution of this work is a GAN-based stain normalization model (SNM) that transfers the staining style of the RC. The proposed system has been designed to maintain morphological features with high visual quality of images and preserve performance when automatically recognized by the accompanying classifier. The overall evaluation has been carried out on three broad classes of cells representative of lymphoma, leukemia and infection, respectively.

The remainder of this section deals with the following aspects: analysis of results, development and training of the model, and computational resource requirements.

7.1. Analysis of results

A key idea in the proposed SNM was the sequential structure with two specialized gain models. GAN1 transforms the initial RGB images into an adaptive gray. The intuitive idea was to fade the image but keeping the morphology as fixed when the blood smear was prepared. The second GAN2 was designed to synthesize stained digital images from scratch while also preserving morphology. For this reason we made the two quantitative comparisons summarized in Table 3 using common metrics in GAN assessments. Test A was designed to evaluate the final output of the SNM via the comparison of the original images from the five centers involved in the study with those digitally stained final images. Test B was designed to have a measure of how the faded images created by GAN1 compared with the same images in grayscale obtained by the classic linear gray transformation. The main conclusion from the FID score comes when observing two facts. First, the values for the RC (3.532 for Test A and 4.418 for Test B) are clearly smaller than the values for the other centers, which is interpreted as the transformations for the RC are practically neutral. Second, differences exist in the FID scores between the four centers, ranging from 9.301 to 29.703 in Test A and from 14.287 to 47.864 in Test B. This can be interpreted in terms of staining variability. The more the digital staining of the centers varies from the RC staining, the higher the score will be. The IS score values are similar for both tests with values between 10.192 and 10.585. This result confirms that the quality of the images with digital staining is similar in the five centers. We consider that this similarity is due to two causes. The first is because the laboratories use the same model of digital morphology analyzer for the acquisition of the images. The second is due to the architecture of the SNM, which is able to manipulate the color spaces without affecting cell morphology. LPIPS values range from 0.039 to 0.068, which are very low values. This indicates that the visual perception to differentiate the staining of the images before and after SNM is poorly discernible, which was a design objective.

In addition to the performance metrics discussed above, an evaluation of the variability of the color channels in the images before and after using the SNM was carried out by using histograms (see Fig. 5). It is remarkable that the histograms of the source images showed high and variable frequencies among the five centers while, after applying the SNM digital stain, the frequencies were reduced and became similar for all centers. We separated the three RGB components and calculated the root mean square error (RMSE) values of the cumulative histograms relative to the RC. It is interesting to analyze these values in Fig. 7. By ordering the RMSE of the centers from smallest to largest they are C3, C2, C4 and C1. The order is the same as for the FID score for Test A from lowest to highest A_{C3}, A_{C2}, A_{C4}, and A_{C1}, discussed above. Consequently, both evaluations (quantitative and visual) indicate the level of digital staining variability between centers adjusted by the SNM.

The third evaluation of the SNM was through automatic recognition of digitally stained cells by a classifier previously trained with images obtained and stained in the RC. First, we compared the performance of the classification of RC images with their original staining (Table 6) versus the classification of SNM-stained images (Table 8). Very slight differences exists between the balanced accuracies in each case: 95.7% and 96%, respectively. Therefore, we can consider that the SNM normalization is neutral when using images from the same center. The next step was a multicenter evaluation, using batch images from four different centers and comparing the performance of the classifier before and after using SNM staining (see Table 9). The first observation is the deterioration of the classification performance when using the original images for the different centers in comparison with the results when using the images from the RC center. While the balanced accuracy was 95.7% for the RC case, the accuracies for C1, C2, C3, C4 before using SNM digital staining were 61.2%, 73.9%, 84.3% and 82.5%, respectively. The second observation to note is that, using the SNM, these accuracies increased to 98.4%, 92%, 93.4% and 92.6%, respectively. By cell class, evidence indicates that the classification of reactive lymphocytes is the most sensitive to staining variability for all centers. When using the original images, the true positive rates (TPR) fell to the range of 46.3% to 66%, and increased considerably to a range between 81.2% and 97.2% when using the SNM stained images. In the case of abnormal lymphocytes, the TPR ranged from 31.9% to 95.7% with the original images and from 83% to 100% with the SNM stained images. For the blast class, the ranges were 90.3% - 94.4% and 94.4% - 100%, respectively. All results are consistent with the general idea that the using the proposed GAN-based normalization staining approach improves the performance of classifiers with multicenter data sets.

7.2. Comparison with other GAN-based approaches

Our objective was to develop a normalization digital staining approach to improve the performance of classifiers of peripheral blood cell images when applied to multicenter datasets after having been trained using images from a reference center. Without any staining normalization, the overall classification average across four external centers using a CNN model trained with RC images was 75.48% (refer to Table 9). Initially, we attempted to improve the accuracy by training the CNN model with the RC database using other color spaces such as HSV, YCbCr and grayscale. We achieved high precision for the RC test set with accuracy values of 94.3%, 93.3%, and 94.7%, respectively for the three-color spaces. However, when testing these models with external centers, we observed a significant decrease in accuracy, with an average global accuracy between centers of only 49.69%, 50.36% and 61.36%, respectively. This was particularly significant for the reactive lymphocyte (RL) group with accuracies lower than 10%. Interestingly, the grayscale space performed better than the other color spaces in terms of accuracy.

We explored an alternative approach by normalizing the images to the RC standard with a single GAN using CycleGAN. Although this architecture combines two distributions to obtain a third, we encountered problems in preserving morphological variability. To address this, we duplicated the original dataset to form a pairwise database. To more accurately simulate staining variations, we generated a database of pairs consisting of RC images and the same images with random digital lighting, color and contrast modifications. This approach was evaluated using the same methodology as in Section 6.2, training the CNN with the RC database and passing the test set through the CNN. However, we observed a decrease in the overall classification results to 86%, with morphological variations visible in the resulting images. When we tested images from external centers using this approach, the overall accuracy was only 50.3%. mWe also noted that the entire RL group was recognized as ALC.

Another attempt was to normalize the images with a single GAN using Pix2Pix. This trial used a pairwise dataset, where the RC images were modified with random transformations of illumination, contrast, and color. The results of this method were promising in the RC test set (98.6% accuracy) and morphological variation was not observed in the resulting images. However, only a slight variation in the color normalization was obtained in the images from external centers compared to the original images. In addition, the average classification accuracy of the external centers only improved slightly to 78.4% compared to images without any normalization (75.48%). The conclusions of this trial were the limitations of Pix2Pix and the need for a better normalization method.

In a subsequent experimentation, we addressed the stain variation by using a single-color component, specifically the gray scale. To achieve this, a Pix2Pix GAN was trained to convert gray images to the RC color space. The database consisted of paired RC images in both RGB and gray formats. The RC test set achieved an overall classification accuracy of 94.1%. In the external centers test, the overall accuracy was 89.26%. This normalization method was considered optimal in terms of computational resources, as it only requires a modification in the color space and a single GAN. However, the reduction in accuracy between RC and the external centers using the normalization system indicated that further improvements were necessary.

This process of testing various approaches ended with the two-stage GANs proposed in this work, the first to normalize to adaptive grayscale and the second to normalize to RC stain. This methodology achieved a classification accuracy of 96% on the RC test set and a global average accuracy of 94.1% on external centers. Overall, this methodology shows promising results in improving classifier accuracy. This system utilizes more computational resources due to the use of two GANs. We discuss the computational requirements of this approach in the following section.

7.3. Computational resources

GANs require a proper training method to create adversity between the generator and the discriminator. The work in [24] uses the CycleGAN architecture as a starting point to change color spaces in tissue samples with simultaneous training of the discriminator and generator. Instead, in this work we used a PIX2PIX architecture to manipulate each input pixel to the system and monitor the result in each training epoch. We opted for this architecture because it learns from a paired database, where color is the variable to modify and cell morphology is the variable to maintain.

The use of PIX2PIX consumes a lot of computational resources for training, so it needs an adequate training technique and optimization of these resources. The learning technique used in GAN 1 and GAN 2 was the NoGAN method [61]. The method trains the generator and the discriminator separately, reducing the computational learning load and avoiding instability in the training losses. The NoGAN technique is widely accessible because it can be used in conjunction with a paired database. The stability of the training allows the correct manipulation of the color space of a cell image without modifying the morphology, avoiding generating an erroneous clinical diagnosis associated with the manipulation of images by the GANs. We conducted 30 epochs of training until the losses decreased considerably (see Fig. 4). At the end of the epochs, the generator and discriminator are trained. However, it is necessary to carry out joint training to adjust the generator weights based on the discriminator prediction. For this reason, we carried out 20 additional epochs of joint training. The overall training strategy was crucial to obtain a correct digital staining

It is interesting to discuss on the training time. Generators G_1 and G_2, and Discriminators D_1 and D_2 completed a separate training epoch in approximately 5, 5.5, 14.5, and 13.5 minutes, respectively. The joint training of $G_1 - D_1$ and $G_2 - D_2$ was approximately 17 and 15 minutes, respectively. The approximate total training time of the SNM was 30 hours.

In addition, we consider important to evaluate the execution time of the entire scheme shown in Fig. 1. That is, pass a set of images through SNM and the classifier, measuring the time to complete the entire process for practical use. We measured the runtime in three tests. Each test used the same set of 1000 images, the initial GPU temperature of 35 degrees Celsius, and 15% of the capacity of the 12 GB Nvidia Titan XP GPU. In Test 1, 1000 images were digitally stained by the SNM. Test 2 consisted in the classification of 1000 images. In Test 3, 1000 images were normalized by SNM and classified. The measured times were 53.43 seconds, 104.09 seconds and 191.82 seconds, respectively.

7.4. Practical use of the system and limitations

In practical use, the proposed system is intended to be used in clinical laboratories as an additional tool to aid in diagnosis, complementing the experience and clinical practice of the pathologist. From a patient's blood sample, the first step in the workflow is smear preparation using staining procedures. The second step is the acquisition of a set of individual cell images by a digital analyzer. In the third step, the pathologist selects a number of images that go through the system in Fig. 1. The expected number of images processed per smear would be less than 100, so the total processing time would be less than 20 seconds in the worst case. The normalization process is done within the SNM and the pathologist only views the original images along with the classification result. The system serves as a support tool to add accuracy and objectivity in case of uncertainty in the morphological inspection, particularly in the presence of pathological cells. It is designed as a complement to pathologists in their role in the diagnosis process.

In this work, the system was evaluated in a specific framework. All training was performed using images from peripheral blood smears collected at the CORE Laboratory of the Hospital Clinic of Barcelona (Spain) over a considerably long period of time, which was the reference center. Smear preparation was performed automatically on the Sysmex SP10 system with the May Grünwald-Giemsa (MGG) staining technique. Cell images were acquired with the Cellavision DM96 digital analyzer. Three cell classes were considered as a prototype case: abnormal lymphocytes, blasts and reactive lymphocytes. They encompass broad subtypes of lymphoma, leukemia and infection, respectively.

With a broad application perspective, this is a limited evaluation. Future work is needed to extend this approach to different image acquisition devices and staining methods. As for digital analyzers, while the Cellavision DM96 has been widely used in many laboratories over the last decade, new devices are entering the market helping to increase the scope of automated digital morphological analysis. To overcome this limitation, transfer learning techniques could be employed when a pre-existing SNM model must be adapted to a new device. In addition, the SNM system was evaluated using MGG staining in samples from four external centers. The results clearly demonstrated that the image classification was significantly improved when compared to the classification of the original images. However, it is important to conduct future studies encompassing a wider range of centers to expand the scope of the normalization approach and enable the exploration of the system's robustness against specific variations in the staining protocol.

While this study considered staining normalization for three cell classes, it is important to note that further opportunities exist to explore the use of SNM in a broader range of classes. For example, the sub-classes of blast cells associated with different types of leukemia and the sub-classes of abnormal lymphocytes are potential targets for research. Furthermore, certain classes of cells may be more prone to changes in staining intensity due to variations in cytoplasmic coloration, making them especially interesting to explore.

8. Conclusions

This work has developed and evaluated a method based on GANs capable of performing normalized digital staining of peripheral blood cell images without modifying their morphology. Two generative adversarial networks have been developed for this system. The first GAN removes staining from the input images, while the second GAN specializes in staining the images with the reference center staining. The user enters the original images with MGG staining through the SNM and at the output synthetic images are obtained with the digital staining of the RC. This staining has been validated by GAN metrics, histogram evaluation, and through artificial classifiers. Synthetic images have been shown to have the same quality as real images. The system requires low computational time in production, little preprocessing time, and allows multicenter adaptability. Therefore, we believe that it can help improve the performance of automatic recognition systems for multicenter final clinical practice.

The main tangible contribution of this work is the methodology and the illustration of its potential in this context. It is interesting to continue working to extend this approach to different image acquisition devices and staining methods, as well as more source centers and more cell classes. As technological advances within this field continue to rapidly evolve, the proposed methodology is subject to improvement. Therefore, future studies can build on this foundation to further improve this methodology in clinical laboratories.

Authors’ contributions

Kevin Barrera developed and training the system, implemented the experiments and contributed to the writing of the manuscript.

JosÃI’ Rodellar supervised the overall project, advised on the deep learning methods, designed the manuscript contents and contributed to writing and editing.

Santiago Alférez contributed to the experiments and their interpretation and contributed to the revision of the manuscript.

Anna Merino supervised the pathology side of the project, advised on the morphological aspects and contributed to the manuscript editing.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work 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.

K. Barrera appreciates the financial support of the Secretary of Higher Education, Science, Technology and Innovation of Ecuador as a personal sponsor entity.

Grants from NVIDIA are appreciated, utilizing donated Titan Xp GPU and instances via SaturnCloud.

All the authors appreciate the collaboration of hematologists and clinical pathologists providing images for this study from the following centers: Hospital German Trias i Pujol, Badalona, Spain; Donostia University Hospital, Donostia, Spain; CellaVision, Lund, Sweden; Hospital Josep Trueta, Girona, Spain.

References

[1] A. Merino, Manual de citología de sangre periférica y líquidos biológicos, Editor. Panamericana (2019).

[2] B. Deepthi, A.K. Prayaga, N. Rukmangadha, et al., Comparison of modified ultrafast giemsa stain with the standard may grunwald giemsa stain in FNAC of various organs, J. Cytol. 39 (4) (2022) 174.

[3] H.O. Lyon, A.P. De Leenheer, R.W. Horobin, W.E.E. Lambert, E.K.W. Schulte, B.M.V. Liedekerke, D.H. Wittekind, Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents, Histochem. J. 26 (2004) 533–544.

[4] N. Michielli, A. Caputo, M. Scotto, A. Mogetta, O.A.M. Pennisi, F. Molinari, D. Balmativola, M. Bosco, A. Gambella, J. Metovic, D. Tota, L. Carpenito, P. Gasparri, M. Salvi, Stain normalization in digital pathology: clinical multicenter evaluation of image quality, J. Pathol. Inform. 13 (2022) 100145, doi:10.1016/J.JPI.2022.100145. /pmc/articles/PMC9577129/ /pmc/articles/PMC9577129/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9577129/.

[5] M. Salvi, A. Caputo, D. Balmativola, M. Scotto, O. Pennisi, N. Michielli, A. Mogetta, F. Molinari, F. Fraggetta, Impact of stain normalization on pathologist assessment of prostate cancer: a comparative study, Cancers 2023, Vol. 15, Page 1503 15 (2023) 1503, doi:10.3390/CANCERS15051503.

[6] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, J. van der Laak, Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology, Med. Image Anal. 58 (2019) 101544.

[7] 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.

[8] L. Boldú, A. Merino, A. Acevedo, A. Molina, J. Rodellar, A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images, Comput. Methods Programs Biomed. 202 (2021) 105999.

[9] A. Acevedo, A. Merino, L. Boldú, A. Molina, S. Alférez, J. Rodellar, A new convolutional neural network predictive model for the automatic recognition of hypogranulated neutrophils in myelodysplastic syndromes, Comput. Biol. Med. 134 (2021) 105020.

[10] A. Molina, J. Rodellar, L. Boldú, A. Acevedo, A. Alférez, A. Merino, Automatic identification of malaria and other red blood cell inclusions using convolutional neural networks, Comput. Biol. Med. 136 (2021) 104680.

[11] H. Eilertsen, C.E. Henriksson, T.-A. Hagve, The use of cellavisionâ c="bad accent" l="1117" n="\v"/>ĉ DM 96 in the verification of the presence of blasts in samples flagged by the sysmex XE-5000, Int. J. Lab. Hematol. 39 (4) (2017) 423–428.

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

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[14] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings (2015), doi:10.48550/arxiv.1511.06434.

[15] M. Mirza, S. Osindero, Conditional generative adversarial nets (2014). doi:10.48550/arxiv.1411.1784.

[16] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN (2017). https://arxiv.org/abs/1701.07875v3.

[17] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired Image-To-Image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, Others, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[20] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[21] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, A. Mukhopadhyay, GANs For medical image analysis, Artif. Intell. Med. 109 (July) (2020) 101938, doi:10.1016/j.artmed.2020.101938.

[22] H. Li, L. Zhang, X. Zhang, M. Zhang, G. Zhu, P. Shen, P. Li, M. Bennamoun, S.A.A. Shah, Color vision deficiency datasets & recoloring evaluation using GANs, Multimed. Tools Appl. 79 (37) (2020) 27583–27614.

[23] Y.-C. Lo, I.-F. Chung, S.-N. Guo, M.-C. Wen, C.-F. Juang, Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application, Appl. Soft Comput. 98 (2021) 106822.

[24] M. Tarek Shaban, C. Baur, N. Navab, S. Albarqouni, StainGAN: stain style transfer for digital histological images, arXiv e-prints (2018) arXiv–1804.

[25] H. Kang, D. Luo, W. Feng, S. Zeng, T. Quan, J. Hu, X. Liu, Stainnet: a fast and robust stain normalization network, Front. Med. (Lausanne) 8 (2021) 2002, doi:10.3389/FMED.2021.746307/BIBTEX.

[26] S. Liu, X. Li, A. Zheng, F. Yang, Y. Liu, T. Guan, Y. He, The generation of virtual immunohistochemical staining images based on an improved cycle-gan, in: International Conference on Machine Learning and Intelligent Communications, Springer, 2021, pp. 137–147.

[27] E. Yuan, J. Suh, Neural stain normalization and unsupervised classification of cell nuclei in histopathological breast cancer images, arXiv preprint arXiv:1811.03815 (2018).

[28] H.R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B.C. Bizzo, Y. Wen, V. Buch, M. Shah, F. Kitamura, M. Mendonça, V. Lavor, A. Harouni, C. Compas, J. Tetreault, P. Dogra, Y. Cheng, S. Erdal, R. White, B. Hashemian, T. Schultz, M. Zhang, A. McCarthy, B.M. Yun, E. Sharaf, K.V. Hoebel, J.B. Patel, B. Chen, S. Ko, E. Leibovitz, E.D. Pisano, L. Coombs, D. Xu, K.J. Dreyer, I. Dayan, R.C. Naidu, M. Flores, D. Rubin, J. Kalpathy-Cramer, Federated learning for breast density classification: A real-world implementation (2020). doi:10.1007/978-3-030-60548-318.

[29] K. Liu, R. Shuai, L. Ma, ZeXu, Cells image generation method based on VAE-SGAN, Procedia Comput. Sci. 183 (2021) 589–595, doi:10.1016/J.PROCS.2021.02.101.

[30] D. Hazra, Y.-C. Byun, W.J. Kim, C.-U. Kang, Synthesis of microscopic cell images obtained from bone marrow aspirate smears through generative adversarial networks, Biology (Basel) 11 (2) (2022) 276.

[31] K. Barrera, A. Merino, A. Molina, J. Rodellar, Automatic generation of artificial images of leukocytes and leukemic cells using generative adversarial networks (syntheticcellGAN), Comput. Methods Programs Biomed. 229 (2022) 107314, doi:10.1016/j.cmpb.2022.107314.

[32] Z. Liu, D. Xu, T. Li, N. Xu, Q. Li, S. Qi, A stain normalization method for the white blood cell image, in: 2019 International Conference on Automation, Computational and Technology Management (ICACTM), IEEE, 2019, pp. 19–23.

[33] A. Tomczak, S. Ilic, G. Marquardt, T. Engel, F. Forster, N. Navab, S. Albarqouni, Multi-task multi-domain learning for digital staining and classification of leukocytes, IEEE Trans. Med. Imag. 40 (2021) 2897–2910.

[34] S. Pipitone, L. Germagnoli, G. Da Rin, A. Di Fabio, A. Fanelli, F. Fiorini, S. Francione, A. Marini, A. Papa, A. Benegiamo, et al., Comparing the performance of three panels rules of blood smear review criteria on an italian multicenter evaluation, Int. J. Lab. Hematol. 39 (6) (2017) 645–652.

[35] E. Gemen, The sysmex SP1000i for automated bone marrow slide smear staining, LABMEDICINE 40 (2009) 23–27.

[36] H. Ceelie, Examen de películas de sangre periférica utilizando microscopía automatizada; evaluación de diffmaster octavia y cellavision DM96, Revista de patología clínica 60 (2007) 35–41.

[37] S. Goel, R.P. Hasserjian, Morphological, flow cytometry, and cytogenetic diagnosis of MDS, in: Diagnosis and Management of Myelodysplastic Syndromes, Springer, 2020, pp. 19–42.

[38] F.J. Reh, Understanding pareto's principle-the 80-20 rule, 2017.

[39] A. Basu, R. Mondal, S. Bhowmik, R. Sarkar, U-Net versus pix2pix: a comparative study on degraded document image binarization, J. Electron. Imag. 29 (6) (2020) 063019.

[40] J. Rodellar, K. Barrera, S. Alférez, L. Boldú, J. Laguna, A. Molina, A. Merino, A deep learning approach for the morphological recognition of reactive lymphocytes in patients with COVID-19 infection, Bioengineering 9 (5) (2022) 229.

[41] J. Ding, B. Chen, H. Liu, M. Huang, Convolutional neural network with data augmentation for SAR target recognition, IEEE Geosci. Remote Sens. Lett. 13 (3) (2016) 364–368.

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 13001–13008.

[43] P. Salehi, A. Chalechale, Pix2pix-based stain-to-stain translation: a solution for robust stain normalization in histopathology images analysis, in: 2020 International Conference on Machine Vision and Image Processing (MVIP), IEEE, 2020, pp. 1–7.

[44] L. Metz, B. Poole, D. Pfau, J. Sohl-Dickstein, Unrolled generative adversarial networks, arXiv preprint arXiv:1611.02163 (2016).

[45] Y. Weng, T. Zhou, Y. Li, X. Qiu, Nas-unet: neural architecture search for medical image segmentation, IEEE Access 7 (2019) 44247–44257.

[46] X. Shao, W. Zhang, Spatchgan: a statistical feature based discriminator for unsupervised image-to-image translation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6546–6555.

[47] 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.

[48] A. Ghosh, H. Kumar, P.S. Sastry, Robust loss functions under label noise for deep neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 24–35.

[49] J.T. Springenberg, Unsupervised and semi-supervised learning with categorical generative adversarial networks, arXiv preprint arXiv:1511.06390 (2015).

[50] L.N. Smith, A disciplined approach to neural network hyper-parameters: part 1–learning rate, batch size, momentum, and weight decay, arXiv preprint arXiv:1803.09820 (2018).

[51] J. Howard, S. Gugger, Deep Learning for Coders with Fastai and PyTorch, O'Reilly Media, 2020.

[52] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).

[53] S. Barratt, R. Sharma, A note on the inception score, arXiv preprint arXiv:1801.01973 (2018).

[54] H. Talebi, P. Milanfar, Learned perceptual image enhancement, in: 2018 IEEE international conference on computational photography (ICCP), IEEE, 2018, pp. 1–13.

[55] M. Stricker, M. Orengo, Similariy of color images, in: Storage and Retrieval for Image and Video Databases III, SPIE, 1995, pp. 1–12, doi:10.1117/12.205308.

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

[57] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, K. Murphy, Progressive neural architecture search, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 19–34.

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

[59] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[60] Z. Zhang, Improved adam optimizer for deep neural networks, in: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), IEEE, 2018, pp. 1–2.

[61] J. Antic, J. Howard, U. Manor, Decrappification, deoldification, and super resolution, in: Facebook f8 conference, 2019, p. 1. https://github.com/jantic/DeOldify.