Computers in Biology and Medicine 131 (2021) 104265

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Improving blood cells classification in peripheral blood smears using enhanced incremental training

Rabiah Al-qudah *, Ching Y. Suen **

Department of Computer Science, Concordia University, 1455 Boulevard de Maisonneuve O, Montréal, QC, Canada

ARTICLE INFO

Keywords:
Automated blood smear analysis
Computer aided diagnosis
Deep learning

ABSTRACT

Peripheral Blood Smear (PBS) analysis is a vital routine test carried out by medical specialists to assess some health aspects of individuals. The automation of blood analysis has attracted the attention of researchers in recent years, as it will not only save time, money and reduce errors, but also protect and save lives of front-line workers, especially during pandemics. In this work, deep neural networks are trained on a synthetic blood smears dataset to classify fifteen different white blood cell and platelet subtypes and morphological abnormalities. For classifying platelets, a hybrid approach of deep learning and image processing techniques is proposed. This approach improved the platelet classification accuracy and macro-average precision from 82.6% to 98.6% and 76.6%–97.6% respectively.

Moreover, for white blood cell classification, a novel scheme for training deep networks is proposed, namely, Enhanced Incremental Training, that automatically recognises and handles classes that confuse and negatively affect neural network predictions. To handle the confusable classes, we also propose a procedure called "training revert". Application of the proposed method has improved the classification accuracy and macro-average precision from 61.5% to 95% and 76.6%–94.27% respectively.

1. Introduction

Blood tests help medical doctors assess patients for certain diseases and conditions, check the function of some body organs and show how well medications and treatments are working. Such tests can only be carried out at specialized medical laboratories by specialists, which expose those specialists to hazards that put them at risk of serious infections. Hence, the automation of blood analysis has attracted the attention of researchers in recent years [1–3], as it will not only save time, money and reduce errors, but also protect and save lives of front-line workers especially during the time of pandemics.

Blood has four main components: Red Blood Cells (RBC), White Blood Cells (WBC), plasma and platelets [4]. RBCs, also known as erythrocytes, carry oxygen and bring carbon dioxide back from the entire body to be released from lungs through breathing. WBCs, also known as leukocytes, are in charge of protecting the body against both foreign invaders and infectious disease. The main types of WBCs are: Neutrophils, Lymphocytes, Monocytes, Eosinophils, and Baso-phils. Platelets, also known as thrombocytes, are responsible for repairing blood vessels in case of injury. Finally, blood Plasma is the liquid part of the blood that carries cells and proteins throughout the body. RBCs are the most common blood cells, for instance, the number of WBCs in adult males ranges from 4.5 to 11.5 thousand in 1 μL, where the number of RBCs in adult males ranges from 4.6 to 6 million in 1 μL [4].

1.1. Peripheral blood smear analysis

Complete Blood Count (CBC) with differential is one of the most commonly requested blood tests in medical labs. It counts RBCs, platelets and the main types of WBCs [5] to evaluate many blood diseases like anemia, and other medical conditions that indirectly affect the blood such as infections. The CBC with differential test is performed by collecting blood in a special tube that prevents clotting, as a first step. Next, the tube is placed in an automated device which analyses and counts the blood components.

In many cases, this test is not enough and more blood analyses are needed and a Peripheral Blood Smear (PBS) is requested. A blood smear, also known as a blood film, is the result of spreading and staining a thin

* Corresponding author.
** Corresponding author.
E-mail addresses: r_alquda@encs.concordia.ca (R. Al-qudah), suen@cse.concordia.ca (C.Y. Suen).

https://doi.org/10.1016/j.compbiomed.2021.104265
Received 2 November 2020; Received in revised form 6 February 2021; Accepted 6 February 2021
Available online 18 February 2021

layer of blood on a microscope slide. Fig. 1 shows the difference between the blood samples needed for each blood test. A blood smear is mainly requested to:

1. Check the presence of immature, and abnormal cells.
2. Identify blood cells that are beyond the capabilities of the automated analyzers, i.e., morphological abnormalities.
3. Verify results obtained by automatic analyzers.

1.2. Related work

The usage of Deep Learning (DL) has attracted the attention of researchers in recent years. It is being widely employed to solve segmentation, classification and detection problems from different application domains [8–10].

Moreover, Haematology and blood smear analysis have been active research topics, that have attracted the attention of people in the medical field [11] and the technology sector [1,12] over the years. Researchers in the IT field have mainly focused on three distinct directions in the context of PBS analysis: Malaria detection [13–16], Leukemia diagnosis [17–20], and blood cell classification [21–26].

In [21,22], the authors classified blood cells into three main categories: RBCs, WBCs, and Platelets. The work of Shahzad and collaborators in 2020 [21] proposed a convolutional encoder-decoder framework with VGG-16 as a feature extraction model. The model achieved an accuracy of 97.45%, 93.34%, and 85.11% for RBCs, WBCs, and platelets, respectively. Whereas in 2018, Evangeline and collaborators [22] proposed a method to segment RBCs from other cells by benefiting from the variation in its intensity compared to other cells' intensities. On the other hand, WBCs and Platelets were segmented based on the area factor. The proposed model achieved an accuracy of 92.5%, 96.7%, and 91% for RBCs, WBCs, and platelets, respectively.

Some other works proposed deeper classification systems that did not only classify the main cell blood types, but also categorised the cells that are classified as WBCs into the five main WBC subtypes, namely, Eosinophil, Lymphocyte, Monocyte, Basophil, and Neutrophil. For instance, Mundhra and collaborators in 2017 [23] utilized 2 separate U-Nets to classify WBCs and Platelets. Next, cells classified as WBCs were further classified into the main WBC subtypes using two CNNs. Moreover, RBCs were classified by applying Otsu's thresholding followed by a CNN that classified the RBCs into 6 subtypes. In Ref. [24] Wang and collaborators proposed an architecture that relied on an ensemble method. In this work, the outputs of n randomly generated CNNs were fused and emitted to an ensemble neural network called PatternNet. The proposed model scored an accuracy of 99.37%.

A handful of papers deepened beyond this level of cell classification, and classified blood cells into its main subtypes and morphological abnormalities. The work of Wang and collaborators in 2019 [25], utilized a private dataset of 14,700 whole-slide images that included 11 categories of WBCs and Platelets. YOLOv3 and SSD networks were trained and tested on different settings to classify the dataset. The YOLOv3 320 × 320 scored a Mean Average Precision (MAP) of 0.92. Moreover, in Ref. [26], Qin and collaborators utilized a private dataset that contained a total of 92,480 WBCs belonging to 40 categories. Many augmentation techniques were applied to handle the dataset imbalance, such as, horizontal and vertical flips, and adding random noises and color changes to the dataset instances. A residual deep neural network consisting of 7 convolutional layers, 2 fully connected layers and three residual blocks was proposed for classification. Seven different schemes were examined by using different activation functions to train the network. The average classification accuracy was 76.84%.

The concept of incremental learning has been proposed in the literature as a way of solving the problem of adapting neural networks to learn to classify instances from classes that were not in the initial training set [27]. For instance, in Ref. [28], a deep network that hierarchically grows to adapt to new classes that were not introduced during the training phase was proposed. In this work, the deep network expanded in a tree-like mode, where classes were categorised, and each category was represented by a branch in the Tree-CNN. Every time a new class was introduced, only the corresponding branch was retrained.

In 2017, Sarwar and et al. [29], introduced a network that incrementally grows by adding convolutional kernels and fully connected layers to the later layers of the network in order to adapt to new classes. An advantage of this method is that the network adapts to new classes without the need to retrain on the previously seen instances, which might not be available at the time of retraining. In Ref. [30], Ferrari and collaborators proposed a method that trains deep neural networks incrementally, using new data and a small sample from the previously seen classes. This method is based on the use of a combination of cross-entropy and distillation loss functions. The works in Refs. [31,32] proposed training schemes that considered exposing the network gradually to the training data in a way that imitates the learning patterns of human beings. The work of Istrate and collaborators in 2017 [31] performed incremental training by first establishing an original network that was later divided into sub-networks. However, this method slightly improved the accuracy of the model.

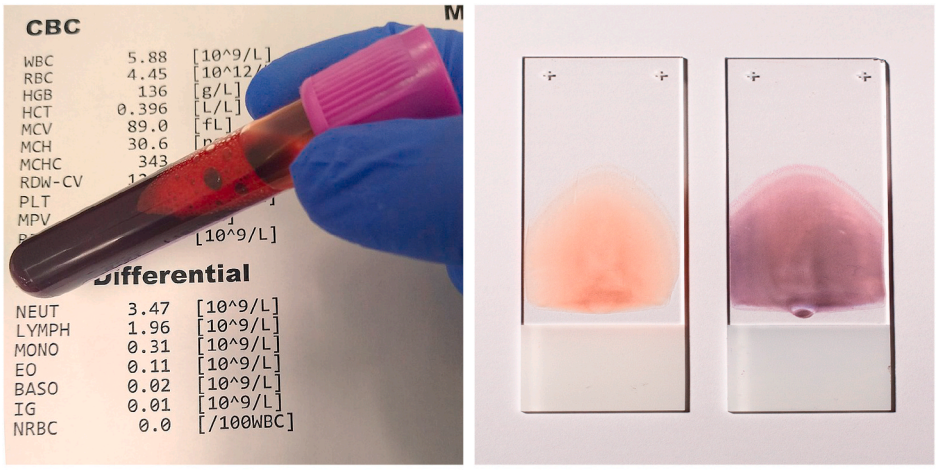


Fig. 1. Different blood samples used in different blood tests.
a: A blood sample collected in a tube for CBC with differential [6]. b: A blood sample spread on a microscope slide for PBS analysis [7].

1.3. Motivation and main contributions

Despite the advancement in haematology, PBS remains as a very important diagnostic test to both haematologists and clinicians; the literature reveals that approximately 70% of clinical decisions are supported by laboratory medicine [33]. Moreover, PBSs play a vital role in the diagnosis and monitoring of disease progression and therapeutic response. For example, PBSs aid in iron-deficiency anemia diagnosis, as RBCs look smaller and paler than normal in patients smears. Moreover, the doctor also requests PBSs periodically to follow up on the anemia progress and measure the effectiveness of the prescribed medication. A well-prepared smear when accompanied with thorough examination and proper interpretation in line with clinical findings can provide a deep assessment of the patients' health.

A lab specialist typically prepares PBS by spreading blood on a clean glass slide, before staining it and setting it to dry. The smear is then examined by an expert under a microscope, hence, PBSs are still being manually analyzed by lab specialists using microscopes. Computer researchers have already developed many methods that automate some aspects of blood analysis [34,35]. Plenty of works have proposed methods that only automate the CBC with differential test, which is already being done using the automatic analyzer in labs. Whereas, only a handful of papers conducted a deeper level of blood analyses to classify the morphological abnormalities and the blood cell subtypes that are beyond the automatic analyzer capabilities. The automation of PBS examination will not only save hematologists time, money and reduce errors, but will also protect and save lives of front-line workers, especially during pandemics.

Table 1 summarizes the main blood cell types and their corresponding subtypes and morphological abnormalities the medical expert examines in PBSs. All the subtypes listed in the table are covered in the datasets utilized for this study, the selection of such types was based on the advice of a medical expert due to their medical significance. Cells classified from the same blood smear are tested among normal ranges in lab, to finally infer whether any of the medical conditions listed in Table 2 [36,37] are present. The table lists 21 general medical conditions caused by the abnormal increase or decrease of WBC and platelet types. Each of those listed medical conditions can reveal the existence of a group of illnesses, for example, Lymphocytopenia when accompanied with a low number of Monocyte indicates that the patient suffers from MonoMAC syndrome. Moreover, monocytosis is seen in chronic bacterial infections, inflammatory conditions such as Crohn's disease, and malignancies such as chronic and acute myeloid leukemia [33]. Hence the automation of the interpretation of PBS can lead to accurate diagnoses of an wide range of illnesses.

This study aims to propose a method for automating PBS analysis using deep learning. Deep learning approaches are data hungry, on the other hand, labeled medical data is scarce [38]. Hence, researchers had to find feasible solutions to increase the size of available medical datasets, like augmentation [39,40]. A shortcoming of augmentation techniques is that it is performed by trial and error, and there is no guarantee that it will enhance results until after training, which might result in repeating the training process [41]. An alternative solution can be creating realistic datasets, i.e. synthetic datasets. In this work the synthetic blood smear dataset from Ref. [42] is utilized for training the neural networks. This synthetic dataset and its annotations were approved by 5 expert hematologists and it provides the following benefits:

1. The synthetic dataset is not subject to any privacy constraints since it does not belong to real people.
2. Rare subtypes, like Plasma cells, are sufficiently present in the dataset.
3. The dataset was automatically annotated for blood cell subtypes and a wide range of morphological abnormalities, without any extra efforts of medical experts.
4. The dataset is balanced in terms of the 15 subtypes covered in this study.

The main contributions of this work are:

1. Propose a novel training method called "Enhanced Incremental Training", that trains the neural network in stages, while assessing confusable classes that negatively affect the network performance. Moreover, to handle the confusable classes, we propose a procedure called "training revert". A set of experiments are presented in this work to verify the effectiveness of the proposed method.
2. Propose a classifier that is trained on a synthetic dataset, and is capable of classifying 15 subtypes and morphological abnormalities. Classifying those classes aids in diagnosing more than 20 medical conditions and diseases. To our knowledge, this is the first classifier designed to work on a synthetic dataset rather than the usual datasets.
3. Train the deep classifier to categorize platelets into 3 main categories, which is rarely discussed in the computer science literature. Moreover, to our best knowledge this is the first work to consider the activated platelets in the classification.

The remaining part of this article is organized as follows. Section 2 describes the datasets utilized in this work. The proposed methods are outlined in Section 3. The results are described in Section 4, the discussion on the results is given in Section 5. Section 6 summarizes the limitation of this work. Finally, section 7 presents the conclusion and future work.

Table 1
Types of peripheral blood cells covered in this study.

WBCs	Atypical Lymphocyte, Atypical Monocyte, Band cell, Basket cell, Basophil, Eosinophil, Hyper-Segmented Neutrophils, Hypo-Segmented Neutrophils, Lymphocyte, Monocyte, Reactive Lymphocyte, Segmented Neutrophils
Platelets	Normal Platelets (Platelets), Giant Platelets, Activated Platelets

Table 2
Medical conditions and disorders the proposed system can infer.

Blood Cell subtype	Abnormal high number	Abnormal low number
Lymphocytes	Lymphocytic, viral infection	Lymphocytopenia
Neutrophils	Neutrophilic leukocytosis, bacterial infection	Neutropenia
Monocytes	Monocytosis, Malaria, fungal infection	
Basophils	Basophilia, malignancy	
Eosinophils	Allergic reaction, fungal infection, Hypereosinophilic syndrome	
Platelet	Thrombocytosis, hypercoagulability	Thrombocytopenia, hypocoagulability
Giant Platelet	Macrothrombocytopenia	
Hypersegmented Neutrophils	Megaloblastic anemia	
Hyposegmented Neutrophils	Pelger-Huët anomaly	

2. Datasets

In this work, a framework is proposed to classify cells that appear on peripheral blood smears into 15 subtypes and morphological abnormalities. The datasets utilized for the purpose of classification are:

1. The synthetic blood smears dataset from Ref. [42] as a training set. The synthetic dataset consists of 2500 whole blood smears and is

annotated for 17 essential categories of blood cells, we will only consider 15 of them in this study. This dataset is balanced and contains around 2000 instances from each class, i.e. 6000 training platelet instances and 24,000 WBC training instances.

2. The ALL-IDB dataset [43] for validation and testing. This dataset contains 39,000 blood elements, but it is only annotated for Leukemia diagnosis purposes. Due to the scarcity of medical data we opted to re-annotate this dataset. A hematologist with 9 years of experience reannotated the dataset to cover the 15 classes considered in this study. The dataset consists of 459 platelet instances and 906 WBC instances. This dataset has been widely utilized in the field of automatic blood cell classification, for example, Ruberto and collaborators [44] utilized this dataset for testing their proposed RBC detection approach and achieved an accuracy of 95.6%. Moreover, the work in Ref. [45] utilized the ALL-IDB dataset for the purpose of leukemia detection with an accuracy score of 94.1%.
3. Blood cell images available on the web: in order to increase the size of the dataset that will be utilized for validation and testing, cell images available on the web were collected. Search keywords like "microscopic band cell" and "microscopic white blood cells" were typed on Google search engine. After a careful observation, images of good visual quality and that were taken in a magnification that falls within the range of 300–500 were considered. One hundred fifty platelet instances and 1100 WBC instances were collected.

The magnification of the cell instances in the three datasets considered in this study range from 300 to 500. Statistics of the datasets are described in Table 3. The All-IDB dataset and the images collected from the web were combined together for validation and testing.

3. Methodology

In image classification applications of DL, a neural network is trained on a set of images until an acceptable error rate is achieved. In the case of supervised learning where all instances are labeled, like in this study, a training dataset of size N can be represented as $\{(x_n, y_n)\}$ for n = 1, …, N, where $x_n \in \mathbb{R}^d$ is the instance feature vector. If there exist M possible output labels, then the labels' set can be expressed as $C = \{c_1, \ldots, c_M\}$, and each instance label which is a subset of the possible M output classes can be expressed as $y_n \in C$. In this work, a set of deep neural networks are utilized to classify 15 blood cell categories, hence, the feature vector of x_n is extracted automatically by the network and the size of the possible output classes set C is 15. When training a classifier we aim to have a discriminant function $f(x \rightarrow \beta_m)$ that maps each instance x_n to M class-specific parameters for each class $\beta_1, \ldots, \beta_M$. Such functions are used to classify each test sample x as the class label that scores the highest β parameter, as shown in equation (1). In this work the softmax function will be utilized to calculate the β values.

$$arg\ max_m[f(x \rightarrow \beta_m)] \quad (1)$$

To tackle the complex nature of blood cell classification problem, the classification process will be divided into 2 parts: the first part deals with platelets classification, and the second part deals with WBCs classification. These approaches are explained in detail in the next two subsections.

Table 3
Dataset statistics.

Cell Type	Train	Validation	Test
Platelet	6000	100	509
WBC	24,000	601	1405

3.1. Classifying platelets

The main challenge in classifying platelets arises from its very small size, which makes it difficult for the network to recognize and extract accurate and distinguishing features [46]. In this work, three categories of platelets are considered; normal platelets (a.k.a platelets), giant platelets, and activated platelets.

The main visual difference between platelets and giant platelets is the length of their diameters. On the other hand, activated platelets look different than the other two categories, because cells that belong to this category have very irregular edges. Pseudopods form on the surface of activated platelets, shaping a star-like appearance once an activation cause is triggered (such as endothelial damage) [47]. Fig. 2 illustrates examples of platelets and activated platelets.

In order to perform this classification task, a one-phase classification technique is enhanced and a two-phase classifier is proposed. In the next two subsections the neural networks that were utilized for classifying platelets and the proposed classification methods will be discussed. Figure 3.1 depicts the steps of both classification approaches.

3.1.1. One-phase classification

The choice of deep classifier can directly affect the accuracy of the classification results, more complex contexts need more complex functions to be learned by the neural network. The literature reveals that deeper networks have performed well in classifying blood cells [48], hence in this work neural networks with a relatively high number of layers will be considered.

For the purpose of classifying platelets the VGG16, VGG-19, ResNet50, InceptionV3, and DenseNet121 networks available on the Keras applications library [49] were experimented, but only the top two networks were considered. All the mentioned networks except the VGG19 and ResNet50 scored validation accuracies less than 70% when trained on the plate-let images. Hence only the VGG19 which consists of 19 layers and the ResNet50 which consists of 50 layers were considered. It is worth mentioning that the reason behind the superiority of the results achieved by the VGG19 and the ResNet networks compared to the results obtained from the other neural networks is considered beyond the scope of this study. Before feeding the images to the VGG19 and the ResNet50 networks, the platelet instances were resized to 200 × 200. Next, the resulting images were augmented by applying rotation, and horizontal and vertical flips. The augmentation techniques and the resizing factor were imperically set. Finally, the images were converted from RGB to BGR, then each color channel was standardized with respect to the ImageNet dataset, without scaling using the Keras "preprocess input" function, the work of M. Shanker and collaborators discusses the benefits of data standardization [50]. Both neural networks were initialized to the ImageNet pretrained weights, and customised by removing the pretrained fully connected layers, and freezing all the remaining layers. Next, an extra fully connected layer and a softmax layer were attached to the networks, finally, the batch size hyperparmeter was tuned and the value 128 was considered. Table 4 summarizes the properties of the fully connected network that was utilized for this task.

3.1.2. Two-phase classification

This approach is proposed to improve the results obtained from the first approach. Here, the same ResNet and a VGG19 networks that were utilized in section 3.1.1 were trained to classify platelets into two classes only; activated platelets and "other platelets". The random undersampling technique was applied in order to reduce the giant and normal platelets training instances to one class while retaining the training dataset balance. Platelets that are classified as "other platelets" are further processed in the second phase by applying a set of image processing techniques. Each instance from the "other platelets" class is first processed by extracting the green channel, because most contours of blood cells appear continuous and contrasted against the background in

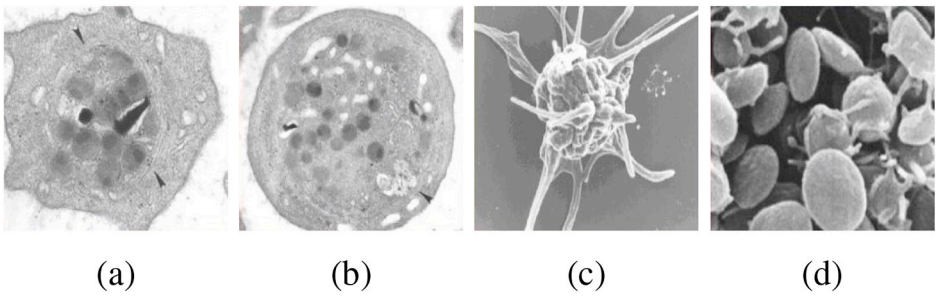


Fig. 2. Activated platelets and normal platelets.
(a) and (c): Activated platelet. (b) and (d): Normal platelet.

Table 4
The properties of the fully connected neural network utilized for platelet and WBC classification.

Framework	Keras 2.4.3
Loss function	Cross Entropy
Learning rate	.01
weight decay	1e-6
Momentum	.9
Optimisation algorithm	Stochastic Gradient Descent
Activation function	ReLU, Softmax
fully connected layer size	1000

the green channel [51]. Next, Contrast Limited Adaptive Histogram Equalization (CLAHE) [52] and Otsu thresholding [53] are applied. All the above mentioned image processing techniques were adopted based on their success in the context of blood cell classification [54–57].

Finally, Hough Circle Transform (HCT) [58] is applied to outline the platelet cell with a circle and measure its diameter. Based on the diameter size, the instance is classified as either normal or giant. This step imitates the way lab technicians classify platelets in a lab. A lab technician typically, estimates the diameter of platelet cells with bare eye under a microscope to decide whether the platelet cell is giant or normal.

3.2. Classifying WBCs

In this work, 12 WBC categories and morphological abnormalities are considered for classification. This relatively high number of classes, in addition to the sensitivity of making wrong classifications in medical diagnoses, adds an extra layer of complexity to this task. Many classes included in this study have high visual resemblance [59–61], which might affect the classification accuracy. Typically, a one-phase classifier is used for this task, however, to accommodate for the complexity of WBCs classification, the one-phase technique is extended to a two-phase classifier:

1. **One-phase WBC classification**: In this approach all WBC instances from all 12 classes are classified by a single neural network. For this task, VGG16, VGG19, ResNet50, InceptionV3, and DenseNet121 networks available on the Keras applications library were experimented, but only the VGG19 was considered for this approach and the enhanced incremental training approach because it achieved the best results. It is worth mentioning that the reason behind the superiority of the results achieved by the VGG19 network compared to the results obtained from the other neural networks is considered beyond the scope of this study. The WBC instances were first resized to 200 × 200. Next, the resulting images were augmented by applying horizontal and vertical flips. The augmentation techniques and the resizing factor were empirically set. Finally, the images were standardized using the Keras "preprocess input" function. The VGG19 network was customised by removing the pretrained fully connected layers, and freezing all the remaining layers. Next, an extra fully connected layer and a softmax layer were attached to the network. The fully connected network utilized in this approach is similar to the one used for platelet classification, which is summarized in Table 4.
2. **Deep neural networks pipeline trained using enhanced incremental training**: In this approach WBCs are classified into 2 phases. In the first phase, all non confusable classes are classified, and classes considered as confusable are passed to the second phase classifiers. For the purpose of figuring out the confusable classes, we opted to propose an enhanced version of incremental deep neural networks training [32], this approach will be explained in detail in this section.

Some causes are extensively discussed in the literature, like class imbalance, which in many studies, has proven its drastic effect on the results., the experiments in this work have shown that some confusable classes not only cause lower correct classification rates in visually similar classes, but also in some other classes that have low visual attributes in common.

3.2.1. Enhanced incremental training

Deep networks are black boxes [62], and knowing what exactly causes wrong predictions can be difficult, or even impossible [63]. Moreover, in any classification task, as the number of classes considered in a dataset increases, it becomes more infeasible to analyze and look for the reason behind high rates of wrong predictions from the confusion matrix. Hence, researchers tend to set some hypothesis on what might be causing the low correct predictions rate, and then handle such hypothesis to measure and test its effect on the neural network performance.

In this work the term **confusable classes** is used frequently. Confusable classes are classes that include instances that have many visual attributes and features in common, which cause a higher level of confusion in the network performance and harms the overall classification results [64,65]. Visual resemblance between different classes can make the classification task more complicated, as feature vectors for their instances come out very similar, which can add more noise to the empirical loss [66] and lead to wrong estimations and results by the function mentioned in equation (1), and hence an increased rate of

wrong classifications. In this study, the number of WBC classes is relatively high, some of these classes are separable and exhibit insignificant visual similarities (e.g., Basket cells and Eosinophils). On the contrary, many other classes are confusable (e.g., Lymphocyte and Atypical Lymphocyte), Fig. 5 shows cells from separable and confusable classes. Hence, Our hypothesis in this study is that the high visual resemblance between some WBC classes is the reason behind the high rate of wrong predictions.

Deep networks are typically trained on instances from all classes at the same time in a single training stage. But in order to test our hypothesis, the incremental training method proposed in Ref. [32] will be enhanced and utilized in a novel way in this work to recognize and handle confusable classes.

In incremental training, a training dataset *TS* is divided into *S* disjoint subsets of size *L*, i.e., each subset contains all the training instances of *L* classes, this implies that:

$$\sum_{i=1}^{S} L_i = |C| \tag{2}$$

where —C— is the size of the set of all possible classes or output labels, S is the number of subsets, and L_i is the number of classes included in the i_{th} subset. The network is initially trained on the first subset, then the resulting weights are passed to the next stage of training, where another subset is added to the training set. A new training process is initiated every time a new subset is added. Eventually, the network training will be completed after *S* training stages. The regime of passing weights between the training stages acts like transfer learning, hence the learned knowledge is accumulated among the stages. We formalised Incremental training in Algorithm 1.

Algorithm 1. Incremental Training

Algorithm 1 Incremental Training

Parameters:

i: training stage index

ts_i: the training set at the ith stage

W_i: the weights resulted from training the network at the ith stage

s_i: the ith subset

$S \leftarrow |C|/L$

$s \leftarrow \{s_0, s_1, ..., s_{S-1}\}$

$W_{-1} \leftarrow InitialWeights$

$ts_{-1} \leftarrow \emptyset$

$i \leftarrow 0$

while $i < S$ **do**

  $ts_i = ts_{i-1} \cup s_i$

  $NeuralNetwork.ReshapeOutput(CurrentShape + L)$

  $W_i \leftarrow NeuralNetwork.Train(ts_i, W_{i-1})$

  $i \leftarrow i + 1$

end while

The proposed enhanced incremental training strategy uti-lises the concept of incremental training as a backbone while adhering additional rules and procedures. In the proposed method, we aim to train the network gradually to identify classes that increase the wrong predictions. Moreover, in order to handle the side effects of classes' visual resemblance, classes are categorised prior to training, and are undersampled into one class during training in case of performance degradation. Undersampling similar classes into one class can ease the network task, as all instances that look similar will be classified the same. After the completion of the proposed enhanced incremental training algorithm, categories that are recognized to include confusable classes will be classified in a second phase. The proposed algorithm is presented in detail in Algorithm 2 and Fig. 4.

Algorithm 2. Enhanced Incremental Training

Algorithm 2 Enhanced Incremental Training

Parameters:
TS: training set
W_i: the weights resulted from training the network at the ith stage
$CategoryTS$: the training set at the beginning of the stage
$Current\_shape$: the number of the network output nodes

▷**Categorization and initialization**

1: $Categories \leftarrow \{Cat_0, Cat_1, Cat_2, Cat_3\}$
2: $W_{-1} \leftarrow ImageNetWeights$
3: $TS \leftarrow \emptyset$
4: $i \leftarrow 0$
▷ **Category looping**
5: **for** $Category$ in Categories **do**
6: $CategoryTS \leftarrow TS$
7: $CategoryOutputShape \leftarrow Current\_Shape$
8: $CategoryWeight \leftarrow W_{i-1}$
9: $TS \leftarrow TS \cup Category$
10: $Nodes \leftarrow Size(Category)$
11: $Network.Reshape(CurrentShape + Nodes)$
12: $W_i \leftarrow Network.Train(TS, W_{i-1})$
▷ **Stage evaluation**
13: **if** $Evaluate(Network) \geq Threshold$ **then**
14: ▷**Training revert**
15: $SubSampledCat \leftarrow SubSample(Category)$
16: $TS \leftarrow CategoryTS \cup SubSampledCategory$
17: $Network.OutputShape \leftarrow CategoryShape + 1$
18: $W_{i-1} \leftarrow CategoryWeight$
19: $W_i \leftarrow Network.Train(TS, W_{i-1})$
20: **end if**
21: $i \leftarrow i + 1$
22: **end for**

The main highlights from Algorithm 2 and Fig. 4 are:

1. **Categorization and initialization**: In line 1 all classes are categorised based on their visual similarity. In this study, the WBC classes are clustered in 4 categories. Table 5 lists each category and its member classes. The Lymphocyte, Monocyte, and Neutrophils category members were grouped based on their high visual similarity, i.e., members of the same group are highly probable to be confused to each other. The Separable WBCs Category comprises the

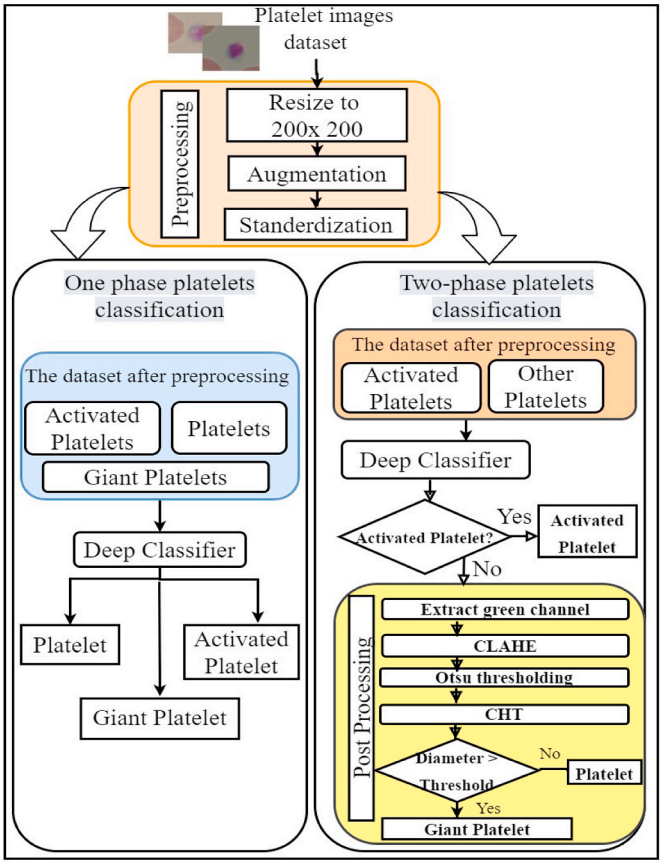


Fig. 3. Platelet classification approaches.

remaining classes that exhibit insignificant morphological and visual similarity and are thought to be easily distinguished from each other. This categorization was planned based on the advice of a medical expert. The categorization was planned based on human's perspective on what classes are confusable, but since neural networks might have a different point of view on what is confusable, the following steps in the algorithm will let the network test the categorization and decide which categories are confusable from the neural network perspective.

2. **Category looping**: In line 5 the algorithm loops on each category, and at the beginning of every training stage, a new category is added to the training set. Next, in lines 6 to 8, the values of all training parameters at the beginning of each stage are kept, to be used if needed later on in running the "training revert" procedure (more details in point 3.2.1). Finally, in lines 9 to 12, a new category is added to the training set while keeping the previously added categories, this act aids in avoiding the "catastrophic forgetting" problem, which refers to the destruction of features learned from previously added data when the neural network is only trained on data of the newly added category [67]. Next, the network output size is adjusted to classify the newly added category, and finally the network weights are assigned to the weights resulted from the previous stage. Utilizing the weights resulted from the previous stages aids in transferring the network knowledge about the previously added categories to the current stage, this way the network will not need to retrain from scratch, but it only fine-tunes the previous knowledge to accommodate the newly added category.
3. **Stage evaluation**: In line 13, the model is tested on the validation set and if the network metrics drop, then we conclude, based on our hypothesis, that the category contains one or more confusable class (es) that caused a higher number of wrong predictions. The threshold of this drop decides whether the category contains confusable classes or not can be numerically or even logically set depending on the problem domain. The criterion that was empirically chosen for this study is 5% drop in the validation accuracy. Hence, the training revert procedure was executed every time the validation accuracy dropped more than 5% compared with the accuracy obtained in the previous stage. It is also worth mentioning that the validation set of WBCs was chosen to contain equal proportions of data from all classes, which facilitated the usage of the accuracy metric. The evaluation of the first model is an exceptional case since there is no previous stage to compare it to, hence, the first model's evaluation accuracy is compared to a baseline value. In this study the baseline value is the accuracy obtained from the one-phase WBC classification as our goal is to improve the results obtained from that approach.
4. **Training revert**: Training the network incrementally facilitates the implementation of gradual testing, which in turn aids in the recognition of the categories that are responsible for the drop in the network performance. Lines 14 to 18, represent the proposed "training revert" procedure, which is only executed in case of encountering a category that contains confusable classes. In this procedure the network reverts to its state before training on the newly added category, i.e., the network reverts to the weights, training set, and output shape values before adding the category that caused the performance drop. Next, in order to handle the complexity associated with the confusable classes, the newly added category is undersampled and reduced to one class. The point of undersampling the category is to retain the training set balance. The undersampled category is then added as one class to the training set. The purpose of adding the category that contains confusable classes as one class is to alleviate the complexity of classifying the confusable classes. As the network will classify all the instances that exhibit a high level of visual resemblance as one class, rather than different classes.

After the completion of the proposed enhanced incremental training algorithm, each category that includes confusable classes will be classified in a separate neural network in a second phase as shown in Fig. 4. Training separate networks to classify categories that contain confusable classes will give the network a chance to better learn and specialize in the features of the confusable classes, and classify them more accurately. In the second phase, each neural network is trained on the corresponding category classes before subsampling.

For the neural networks of the second phase, multiple deep networks were experimented but the VGG16 was considered because it achieved the best results. The VGG16 classifier was connected to a shallow neural network that consists of a fully connected layer, a dropout layer and finally a softmax layer. The cross entropy loss was used for this task and the batch size, image resolution, and dropout rate hyperparameters were tuned for each of the second phase networks separately. Multiple combinations of augmentation techniques are also experimented for each of the networks to improve the results. The early stopping technique was utilized to stop the training and avoid overfitting.

3.2.2. *Batch size tuning*

A key factor in improving the classification results in the context of peripheral blood cell classification is tuning the batch size hyperparameter [25]. A neural network trains with weights W can be seen as an optimisation problem, where we aim to adjust W to reach the minimum possible value for the loss function F. This optimisation problem can be expressed as:

$$arg\ min_W F \tag{3}$$

In this work the mini-batch variant of Stochastic Gradient Descent (SGD) algorithm is used to solve the optimisation problem in equation (3). During training, at the j-th iteration, W is updated using the formula:

$$W_{j+1} = W_j - \alpha/mb\Delta W_j \tag{4}$$

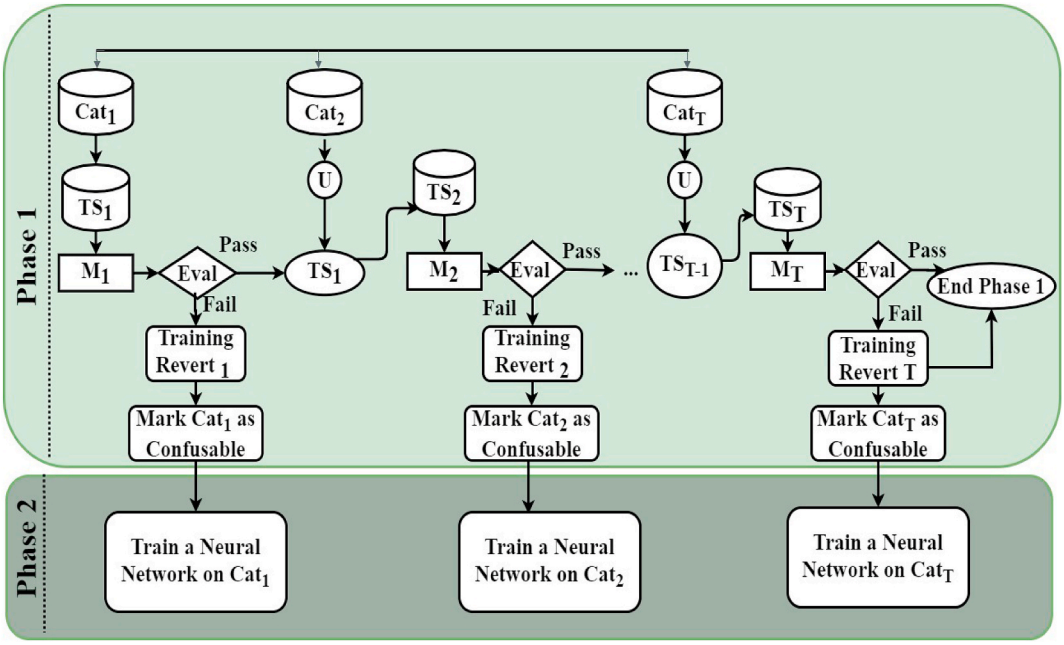


Fig. 4. The enhanced incremental training framework.

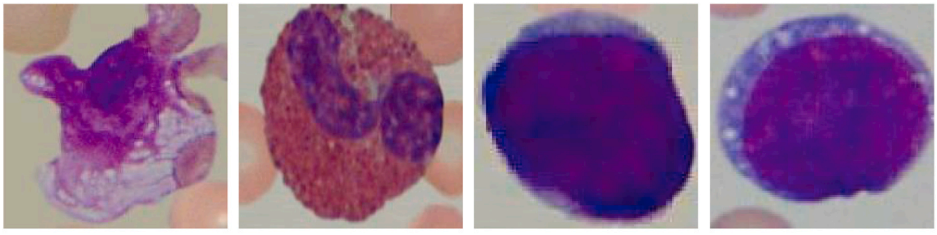


Fig. 5. Instances from separable and visually similar blood cell types, with microscope at magnification 300×.
(1) Separable instances: a: Basket cell b: Eosinophil (2)Confusable instances c: Lymphocyte d: Atypical Lymphocyte [43].

where α is the learning rate and *mb* is the batch size [68]. It can be seen from equation (4) that the batch size can directly affect the value of the updated weights, and subsequently affect the value of the loss *F* that we aim to minimize. In order to tune the neural networks used for WBC classification, the deep networks trained using the enhanced incremental training method will be run on 4 commonly used batch size values, to examine the effect of tuning this hyperparameter on the classifiers' predictions. Four widely used batch size values were chosen for the experiments; 16, 32, 64, and 128. The value 128 was the highest possible value given the available resources.

3.3. Considering determinism

Neural networks use randomness by design. Many forms of randomness are utilized during training, such as, random weights initialization, random mini-batching, random augmentation. This randomness implies that different results will be obtained when training the exact same neural network on the exact same training data multiple times. Introducing determinism to deep networks means controlling the random processes to generate the same random numbers every time in order to guarantee reproducibility.

Since the proposed "enhanced incremental training" method includes training neural networks gradually, while assessing the difference between the evaluation metrics between stages, considering training the neural networks while taking determinism into account is crucial. This is simply because if the network was non-deterministic, then a different neural network with different parameters will be produced at every stage.

In order to enforce determinism on the training process of all the classifiers considered in this work, technical solutions proposed by Keras [69] and Nvidia [70] were followed. In addition, all neural networks were initialized with weights pretrained on the ImageNet dataset, which eliminated the randomness caused by the weight initialization factor. To the best of our knowledge, this is the first work to consider determinism with incremental training.

Table 5
WBC categories.

Category ID	Category Name	Category Members
Cat_0	Separable WBCs Category	Band cell, Basket cell, Basophil, Eosinophil
Cat_1	Monocyte Category	Monocyte, Atypical Monocyte
Cat_2	Lymphocyte Category	Atypical Lymphocyte, Reactive Lymphocyte, Lymphocyte
Cat_3	Neutrophils Category	Hyper-Segmented Neutrophils, Hypo-Segmented Neutrophils, Segmented Neutrophils

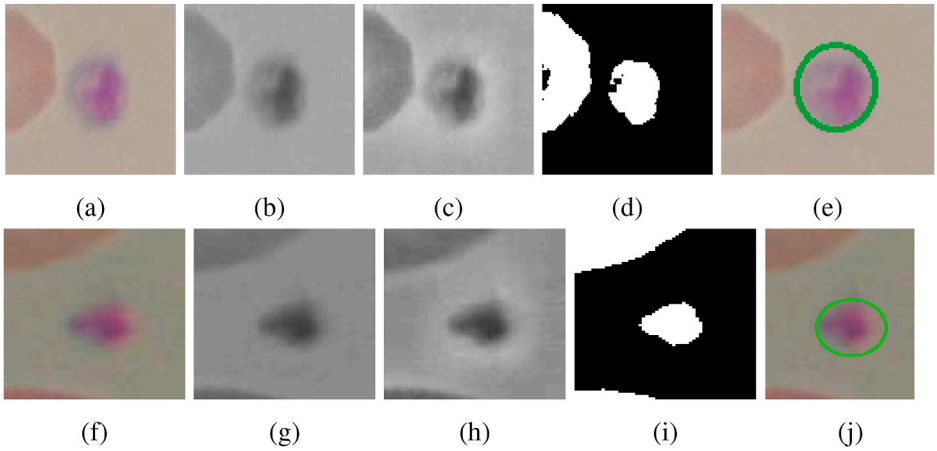


Fig. 6. Sample results from the second phase of classification.
(a) to (e): test instance 1 classification process, (f) to (j): test instance 2 classification process(a),(f): the original image; (b),(g): the extracted green channel; (c),(h): the resulting image after applying CLAHE; (d),(i): the resulting image after applying Otsu's thresholding; (e),(j) the platelet is outlined after applying Hough Circle Transform.

3.4. Evaluation

In order to evaluate the platelet and WBC classification approaches, cross-validation is used. As mentioned in section 2, the datasets were partitioned into training, validation, and testing sets. Since determinism is considered in this work, the experiments will not be repeated, as the results will not change.

The final models that result from both platelet classification approaches, and both WBC classification approaches will be tested on the test set and evaluated using the precision, recall of each class. Moreover, each model will be evaluated in terms of the macro-average precision, the macro-average recall, and the accuracy [71].

4. Results

In this section the results of classifying platelets and WBCs using the approaches listed in the methodology section will be presented.

4.1. Platelets classification results

Two approaches are proposed in this work for platelets classification; one-phase classification, and two-phase classification. The results of both approaches, are presented in the next two sub sections.

4.1.1. One-phase classification

In this approach, a single neural network was trained to classify all 3 platelet types. A VGG19 network was trained first, but it only achieved 78% test accuracy. In the second experiment, a ResNet50 was trained to classify all 3 classes. The trained ResNet50 achieved a test accuracy score of 82.67%.

4.1.2. Two-phase classification

In this approach platelets are classified in two phases. In the first phase, a VGG19 network was trained and tested on both classes. This network achieved an accuracy score of 98.4%. Classifying both confusable classes (normal and giant platelets) as one class reduced the level of confusion, and the rate of misclassifications and resulted in a better accuracy.

In the second experiment, a ResNet50 was also trained and tested on both classes. This network test result surpassed all the previous networks trained in this study for platelet classification. An accuracy test score of 100% was achieved by this network, hence, we chose to only consider it for the second phase. In the second phase, the instances classified as "other platelets" went through a set of image processing techniques and filters, as illustrated in Fig. 3.

Fig. 6 illustrates the classification process of two test samples in the second phase, each row in the figure represents one test sample. Only a few instances were misclassified in the second phase, leading to an overall accuracy score of 98.6%.

Table 6
One phase WBC classification results.

Batch Size	Accuracy (%)
16	33.7
32	39.1
64	44
128	61.5

4.2. WBC classification results

Two approaches are proposed in this work for WBC classification; one phase WBC classification, and deep neural networks pipeline based on enhanced incremental training.

4.2.1. One phase WBC classification

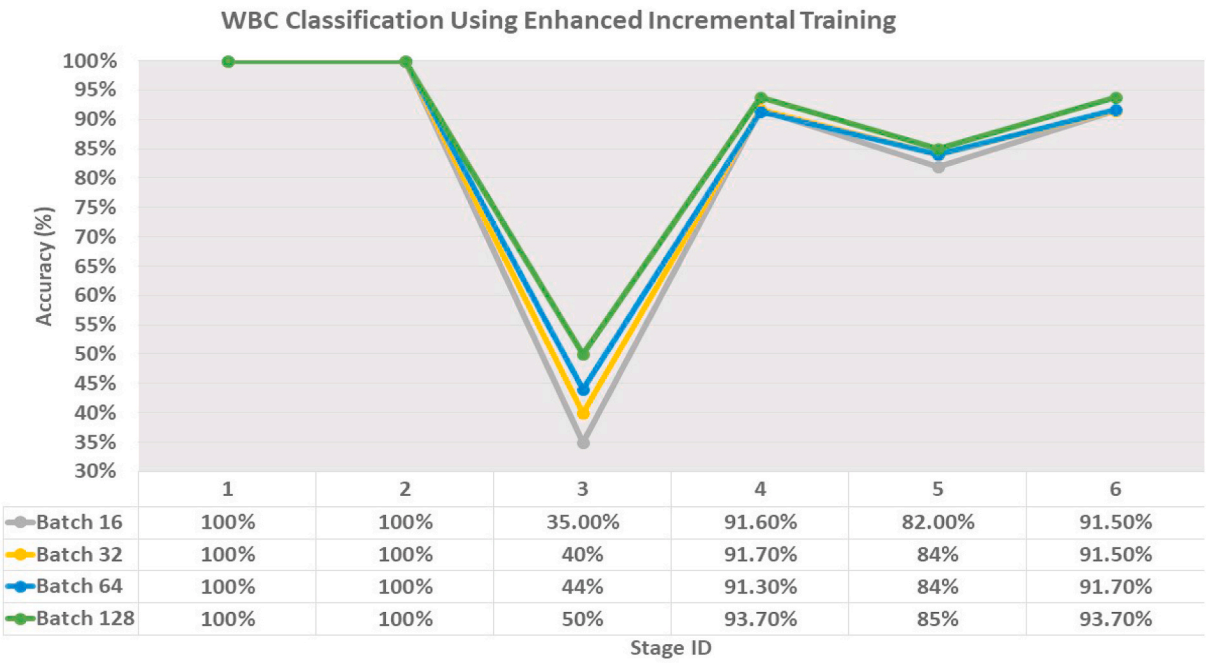
In this approach all WBC instances from all 12 classes are classified by a single VGG19 neural network.

In the second approach of WBC classification, experiments are run with four different batch sizes. Hence, in order to make the results obtained from this approach comparable to the ones from the second approach, the VGG19 network is trained four times with each of the four batch size values. The VGG19 network was customised by removing the fully connected layers, and freezing all the remaining layers. Next, an extra fully connected layer and a softmax layer were added to the network.

The results obtained are listed in Table 6. The poorest result was

Table 7
Classes included in each stage.

		Stage ID					
Category	Class Name	1	2	3	4	5	6
Cat_0	Band cell	✓	✓	✓	✓	✓	✓
Cat_0	Basket cell	✓	✓	✓	✓	✓	✓
Cat_0	Basophil	✓	✓	✓	✓	✓	✓
Cat_0	Eosinophil	✓	✓	✓	✓	✓	✓
Cat_1	Monocyte		✓	✓	✓	✓	✓
Cat_1	Atypical Monocyte		✓	✓	✓	✓	✓
Cat_2	Reactive Lymphocyte			✓		✓	✓
Cat_2	Atypical Lymphocyte			✓	✓		
Cat_2	Lymphocyte			✓			
Cat_3	Hyper-Segmented Neutrophils					✓	✓
Cat_3	Hypo-Segmented Neutrophils					✓	
Cat_3	Segmented Neutrophils					✓	



	1	2	3	4	5	6
Batch 16	100%	100%	35.00%	91.60%	82.00%	91.50%
Batch 32	100%	100%	40%	91.70%	84%	91.50%
Batch 64	100%	100%	44%	91.30%	84%	91.70%
Batch 128	100%	100%	50%	93.70%	85%	93.70%

Fig. 7. The workflow of training the four neural networks using enhanced incremental training and the accuracy results obtained from each stage using the validation set.

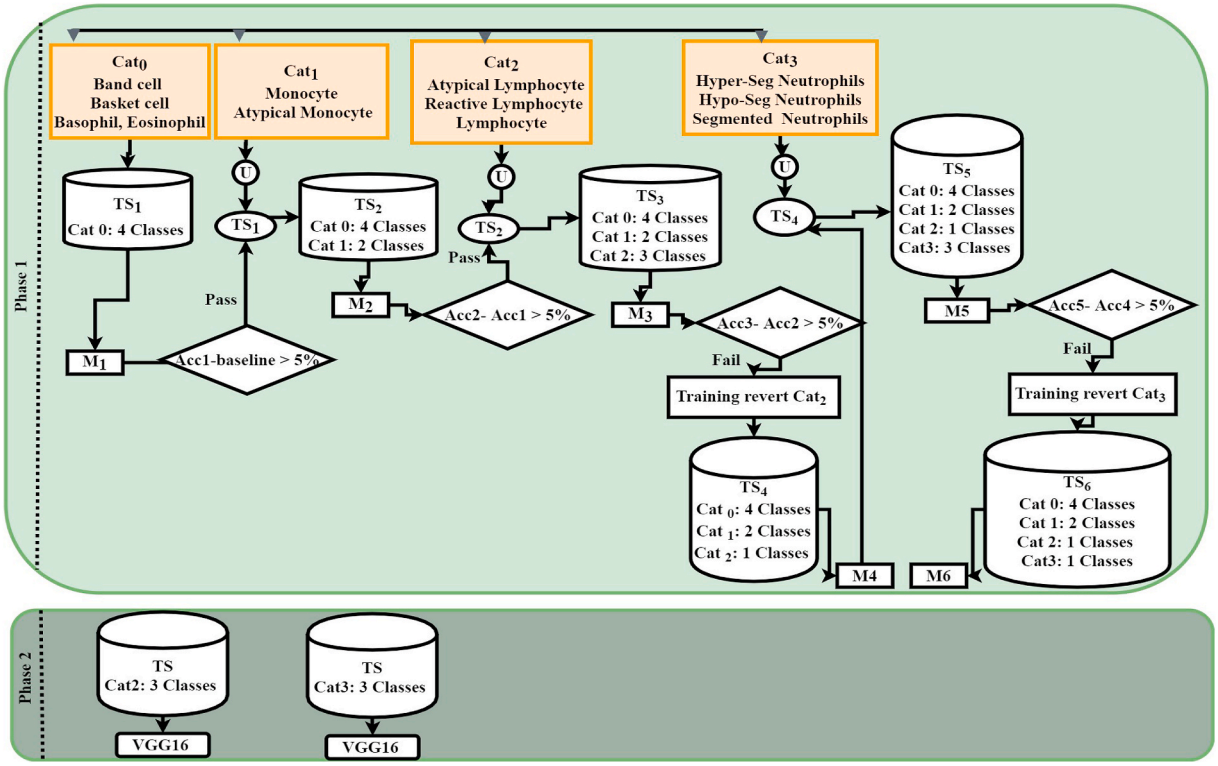


Fig. 8. The workflow of training using the proposed training scheme.

obtained with the smallest batch size. On the other hand, the results kept improving as the batch size value increased. However, the best achieved result was 61.5%.

4.2.2. *Deep neural networks pipeline based on enhanced incremental training*

In this approach WBCs are classified into two phases. During the implementation of the proposed method, Cat_0 was added to the training set first, as according to the medical expert, the odds of having a confusable class in this category is very low. Hence, training this category first helps to have a solid basis of trained weights before training the network on the rest of the categories which contain classes with high visual resemblance.

After the completion of each stage, the network was evaluated based

on an evaluation criterion in order to decide if the category that was added at the beginning of the same stage has negatively affected the network generalizability (see line LABEL:criterion in Algorithm 2). The Criterion that was empirically chosen for this study is 5% drop in the validation accuracy. Hence, the training revert procedure was executed every time the validation accuracy dropped more than 5% compared to the accuracy obtained in the previous stage. In the proposed method, the network is tested on the validation set during training and only the final model that results from training the network on all categories is tested on the test set. Four neural networks were trained using enhanced incremental training, each classifier only differed in the batch size value.

Table 7 lists the classes that were included in each stage. Fig. 8 depicts all training stages for each of the four neural networks, along with the validation accuracy results obtained from each stage. It is clear from the figure that all classifiers followed the same pattern during the training process. Moreover, Fig. 7 depicts the workflow of executing the proposed training scheme. Some highlights from Table 7, Figs. 7 and 8 regarding all four classifiers are:

- The validation accuracy was 100% after adding Cat_0 and Cat_1. Despite our expectations that the accuracy will drop after adding Cat_1 due to the visual similarity of the Monocyte and atypical Monocyte cells, the model actually kept a high accuracy score.
- In Stage 3, classes of category Cat_2 were added and the accuracy score dropped by more than 5%. Hence, Cat_2 was identified as a category that comprises confusable classes. The confusion matrix at this point was analyzed, and it was observed that the drop in the accuracy was mainly due to the high level of confusion between the instances of the three lymphocyte category classes, which supports the assumption that visual resemblance can be a key factor in negatively affecting the classification results.
- In Stage 4, the "training revert" procedure was executed; the weights and the training set were reverted to their values just after training category Cat_1, i.e., Stage 2. All Cat_2 classes were then randomly undersampled and reduced to one class. The network was trained again with the reduced category and the accuracy score increased.
- The accuracy dropped again by more than 5% after adding the last category, i.e. Cat_3. Hence, the "training revert" procedure was executed again, and the entire category was reduced to one class and added to the training set. The accuracy score increased and the final validation accuracy was higher than 90% for all batch sizes.
- All classifiers that were trained on different batch sizes showed the same pattern during the training stages, hence the observations above apply to all four classifiers.
- Comparing the accuracies of the four classifiers that were trained on different batch sizes, it is noted that increasing the batch size increased the accuracy results.

All the results shown in Fig. 7 are the validation set results, the final model resulted from Stage 6 is the only one that was tested on the test set. The results of the test set were very similar to the validation results. The highest test accuracy result obtained was 95% achieved by setting the batch size value to 128. Table 8 lists all test accuracy results.

In the second phase to classify Cat_2 and Cat_3, two VGG16 networks were utilized; the first one was trained to classify the Lymphocyte category into, Lymphocyte, reactive Lymphocyte, and atypical Lymphocyte. The second network was trained to classify the Neutrophils category into hypo-segmented Neutrophils, hyper-segmented Neutrophils, and segmented Neutrophils.

Table 8
Test results of the WBC classifiers trained using the proposed method.

Batch Size	Accuracy (%)
16	91
32	91.3
64	92.2
128	95

Table 9
Comparison of the results of WBC and platelet classification using our proposed system with existing works in the literature in terms of Accuracy.

Method	Accuracy (%)
Our proposed work (WBCs & platelets)	96.7 (CI 95% 96.69,96.709)
Wang et al. [25] (SSD 300 × 300)	75.1 (CI 95% 72.1,78.1)
Qin et al. [26] (shallow residual network)	60 (CI 95% 57.3,62.7)

The best Lymphocyte classification results were obtained with a batch size value of 128. Moreover, the literature reveals that the size of the input image of a neural network aids in improving its performance [72,73]. Hence, 10 different resolutions ranging from 200 to 600 were experimented: {200, 250, 300, …, 600}, where 200 is the original image size emitted from the main classifier in phase 1, and 600 is the highest value possible with the available hardware. The lowest accuracy was 92%, scored with the lowest resolution. Moreover, the accuracy kept improving as the input resolution value increased. Hence, all instances were upscaled to 600 × 600.

On the other hand, to classify the Neutrophils category, another VGG16 classifier was tuned to upscale all instances to 400 × 400, this input resolution was tuned using the same procedure followed in the Lymphocyte category classifier. The lowest accuracy score was 95%, achieved with the lowest resolution, and the accuracy score was 100% for all resolution values higher than 350. It is worth mentioning that only the best classifier from phase one was tested with the classifiers from phase two, and both networks in phase two achieved 100% accuracy.

To sum up, during the test phase every test instance will be first fed to the main VGG19 neural network trained in the first phase, the instances that are classified as either Lymphocyte category or Neutrophil category are further emitted to the networks in the second phase for more detailed classification.

4.3. Comparison with existing works

A comparison of our proposed method with the works of Wang et al. [25] and the work of Qin et al. [26] is shown in Table 9. In order to do the comparison, the overall accuracy score of the top platelet and WBC classification models was considered. Next, the entire training set (i.e. platelets and WBCs) was fed to the networks reproduced from both [25,26]. The results in the table show the superiority of our proposed model. It is also noted that the results of the models from the literature were relatively close to the accuracy scores obtained from implementing the one-phase platelet classification and one phase WBC classification.

The method proposed in Ref. [25] scored an overall accuracy of 90.09% in the original paper, but it only achieved an accuracy score of 75.1% when trained and tested on the datasets of this study. One possible reason for this is that the method in the original paper was trained on a dataset that was only annotated for 11 types of platelets and WBCs, and this combination did not comprise many confusable classes

Table 10
Confusion matrix resulted from the one-phase classification approach.

		Predicted		
	Activated	Giant	Normal	Recall
Activated	50%	50%	0%	50%
Giant	0%	76%	24%	76%
Normal	1%	16%	83%	83%
Precision	98%	54%	78%	
Accuracy	82.67% (CI 95% 79.11, 86.23)			
Macro-average recall		69.6%		
Macro-average precision		76.6%		

Table 11
Confusion matrix resulted from the two-phase classification approach.

		Predicted		
	Activated	Giant	Normal	Recall
Activated	100%	0%	0%	100%
Giant	0%	94%	6%	94%
Normal	0%	1%	99%	99%
Precision	100%	99%	94%	
Accuracy	98.6% (CI 95% 97.5, 99.7)			
Macro-average recall		97.6%		
Macro-average precision		97.6%		

which made the classification task easier. Moreover, the method reproduced from Ref. [26] scored the lowest accuracy, possibly because of its shallow structure which did not comply to the level of the problem complexity.

5. Discussion on the results

In this section the results of the top models obtained from both approaches will be discussed in detail.

	Predicted												
	A.Lymphocyte	A.Monocyte	Band	Basket	Basophil	Eosinophil	Hyper S Neutrophil	Hypo S Neutrophil	Lymphocyte	Monocyte	R. Lymphocyte	S.Neutrophil	Recall
A. Lymphocyte	57%	0	0	0	0	0	0	14%	15%	14%	0	0	57%
A.Monocyte	0	100%	0	0	0	0	0	0	0	0	0	0	100%
Band Cell	0	0	100	0	0	0	0	0	0	0	0	0	100%
Basket Cell	2%	0	0	94%	0	0	0	0	4%	0	0	0	94%
Basophil	0	0	0	0	100%	0	0	0	0	0	0	0	100%
Eosinophil	0	0	0	0	0	100%	0	0	0	0	0	0	100%
Hyper S Neutrophils	0	0	0	0	0	0	86%	0	0	0	0	14%	86%
Hypo S Neutrophils	0	0	0	0	0	0	0	100%	0	0	0	0	100%
Lymphocyte	37%	0	0	7%	0	0	0	1%	53%	1%	1%	0	53%
Monocyte	0	0	0	0	0	0	12.5%	0	0	75%	0	12.5%	75%
R. Lymphocyte	17%	0	0	0	0	0	0	0	66%	0	17%	0	17%
S.Neutrophils	0	0	0	9%	0	0	34%	0	0	0	0	57%	57%
Precision	50%	100%	100%	85%	100%	100%	65%	87%	38%	83%	94%	68%	
Accuracy	**61.5%**	Macro-average recall			**78.25%**			Macro-average precision			**76.6%**		

Fig. 9. The normalized confusion matrix of the top model trained using the one phase WBC classification method.

	Predicted								
	A.Monocyte	Band Cell	Basket Cell	Basophil	Eosinophil	Lymphocyte Cat	Monocyte	Neutrophils Cat	Recall
A.Monocyte	100%	0	0	0	0	0	0	0	100%
Band Cell	0	100%	0	0	0	0	0	0	100%
Basket Cell	0	0	95%	0	0	5%	0	0	95%
Basophil	0	0	0	100%	0	0	0	0	100%
Eosinophil	0	0	0	0	100%	0	0	0	100%
Lymphocyte Cat	0	0	4%	0	0	96%	0	0	96%
Monocyte	0	0	0	0	0	12%	88%	0	88%
Neutrophils Cat	0	4%	5%	0	0	0	0	91%	91%
Precision	100%	71.43%	84%	100%	100%	98.77%	100%	100%	
Accuracy	**95%**	Macro-average recall		**96.25%**		Macro-average precision		**94.27%**	

Fig. 10. The normalized confusion matrix of the top model trained using the proposed method.

5.1. Platelet classification results

Two platelet classification approaches were experimented in this work, Table 10 shows the confusion matrix of the one-phase ResNet50 model, and Table 11 shows the confusion matrix of the two-phase ResNet50. As can be seen in Table 10, one reason for the low accuracy in the one-phase approach was the high visual similarity between giant and normal platelets. For instance, 24% of the giant platelets were confused with platelets, and 16% of the platelets were confused with giant platelets, which negatively affected the overall classification accuracy. Hence, combining both classes in the "other platelets" class aided in improving the classification results. It is also worth mentioning that the activated platelets classification results were also another reason for the low accuracy of the one phase approach but investigating the reason is beyond the scope of this study, as this class is not considered as confusable, hence it is not related to the study hypothesis. On the contrary, in the two-phase approach only 6% of the giant platelets were confused with platelets, and 1% of the platelets were confused with the giant platelets.

5.2. Discussion on the WBC classification results

In the proposed method, classes that are thought to be confusable were categorised based on human's perspective, One option to implement the classification task could be subsampling all confusable categories; train neural networks in the second phase, but since neural networks might have a different point of view on what is confusable, the next steps in the algorithm will let the network test the categorization and decide which categories are confusable from the neural network perspective. In this section the results of the top models obtained from both WBC classification approaches will be discussed. Fig. 9 shows the 12×12 normalized confusion matrix of the top model trained using the one phase approach when tested on the test set, and Fig. 10 shows the 8×8 normalized confusion matrix of the top model trained using the proposed approach when applied to the test set. The latter one is smaller in size because two out of the four categories were subsampled during training.

The following insights can be observed about each of the four categories from the confusion matrices:

1. Classes of Cat_0:
   (a) The Recall values for the Band cells, Basophils, Eosinophils and Basket cells were approximately the same in both experiments.
   (b) The precision values for the Basophils, Eosinophils and Basket cells were approximately the same in both experiments. It is also noted that the precision score of the Band cells decreased in the model trained with the proposed method.
2. Classes of Cat_1:
   (a) The recall score of the Atypical Monocyte class remained 100% in both experiments, whereas, the proposed method improved the recall of the Monocyte class from 75% to 88%.
   (b) The precision score of the Atypical Monocyte class remained 100% in both experiments. On the other hand, the proposed method improved the Monocyte class precision from 83% to 100%.
3. Classes of Cat_2: There is no exact class-to-class comparison for this category because the classes of this category were blended into one class during the execution of the proposed method. The classes of this category were highly confused with each other in the one-phase approach, for example:
   (a) 37% of the Lymphocytes were incorrectly classified as atypical ones.
   (b) 15% of the atypical Lymphocytes were misclassified as Lymphocytes.
   (c) 17% of the reactive Lymphocytes were incorrectly classified as atypical Lymphocytes. Moreover, 66% of the reactive Lymphocytes were misclassified as Lymphocytes.

The recall and precision ratios of the Lymphocyte category in the proposed method were 96%, and 98.7% respectively.

4. Classes of Cat_3: The following observations can be made from the one phase approach:
   (a) 14%of the hyper-segmented Neutrophils were incorrectly classified as segmented Neutrophils.
   (b) 34% of the segmented Neutrophils were misclassified as hyper-segmented Neutrophils.
5. The accuracy, macro average recall, and the macro average precision of the proposed method are 95%, 96.25%, and 94.27% respectively. Which is higher than the results obtained from the one phase approach.

After the proposed system finishes classifying the cells in a blood smear, cells are grouped based on their types and then counted. Next, those counts are compared against normal ranges in order to produce the test results by reporting which of the clinical conditions listed in Table 2 are present. Overall, it can be seen that the proposed method has improved the quality of the classification results. A drawback of utilizing the proposed method is that it needed 3 neural networks to complete the task, compared to one neural network in the one phase approach. But it is worth mentioning that automatic PBS analysis is not a real time application, and like other medical applications, the correctness of the results is a high priority.

The proposed method can be utilized in other application domains where confusable classes are thought to be the reason behind low classification results. Moreover, the proposed method was utilized on top of pretrained models available in an open source framework i.e. Keras. This indicates that the method can be easily reproduced and deployed by other researchers.

6. Limitations

The authors do not have access to values of human-level performance. Hence, only computer based performance is reported.

7. Conclusion and future work

Blood cell classification has been an active research topic in recent years. While many studies have proposed classifiers for the main blood cell types, only a handful of papers conducted a deeper level of blood analyses to classify the morphological abnormalities and the blood cell subtypes that are beyond the automatic analyzer capabilities. In this work, deep classifiers are trained to identify 15 WBC and platelet subtypes and morphological abnormalities of blood cells with a dataset of synthetic blood smears.

For classifying platelets, a one-phase classification approach is experimented and a hybrid approach of deep learning and image processing techniques is proposed. The proposed approach handled the confusion caused by the visual resemblance between the giant and normal platelets and improved the platelet classification accuracy and macro-average precision from 82.6% to 98.6% and 76.6%–97.6% respectively.

Moreover, for white blood cell classification, a novel sch-eme for training deep networks is proposed, namely, Enhanced Incremental Training, that automatically recognises and handles classes that confuse and negatively affect neural network predictions. To handle the confusable classes, a procedure called "training revert" is also proposed. Application of the proposed method has improved the classification accuracy and macro-average precision from 61.5% to 95% and 76.6%–94.27%. The proposed method can be utilized in other application

domains where confusable classes are thought to be the reason behind low classification results. Moreover, the proposed method was utilized on top of pretrained models available in an open source framework i.e. Keras. This indicates that the method can be easily reproduced and deployed by other researchers.

For future work, we plan to test the enhanced incremental training method on large-scale medical and non-medical datasets. Moreover, we plan to examine the influence of some factors on the results, such as, category orderings, and the usage of more complicated criteria to evaluate the training progress between stages. Finally, we plan to experiment with other validation strategies such as the K-fold cross validation technique.

Declaration of competing interest

None

References

[1] Cecilia Di Ruberto, Andrea Loddo, Lorenzo Putzu, Detection of red and white blood cells from microscopic blood images using a region proposal approach, Comput. Biol. Med. 116 (2020) 103530.

[2] M. Roy Reena, P.M. Ameer, Localization and recognition of leukocytes in peripheral blood: a deep learning approach, Comput. Biol. Med. 126 (2020) 104034.

[3] Satishkumar L. Varma, Satishkumar S. Chavan, Detection of malaria parasite based on thick and thin blood smear images using local binary pattern, in: Brijesh Iyer, S. L. Nalbalwar, Nagendra Prasad Pathak (Eds.), Computing, Communication and Signal Processing, Springer Singapore, Singapore, 2019, pp. 967–975.

[4] Bernadette F. Rodak, A. George, Fritsma, Doig Kathryn, Hematology: Clinical Principles And Applications, Elsevier Health Sciences, 2007.

[5] Beverly George-Gay, Katherine Parker, Understanding the complete blood count with differential, J. PeriAnesthesia Nurs. 18 (2) (2003) 96–117.

[6] Wikipedia, A cbc specimen in front of a printout displaying cbc and differential results. https://en.wikipedia.org/wiki/Complete_blood_count, 2020.

[7] Cc By-Sa 3, 0 Coinmac Own work. Stained and unstained blood smears. https://commons.wikimedia.org/w/index.php?curid=20155466, 2020.

[8] Mrouj Almuhajri, Ching Suen, Intensive survey about road traffic signs preprocessing, detection and recognition, in: Advances in Data Science, Cyber Security and IT Applications, Springer International Publishing, Switzerland, 2019, pp. 275–289.

[9] Ali K. Hmood, Ching Y. Suen, An Ensemble of Character Features and Fine-Tuned Convolutional Neural Network for Spurious Coin Detection, vol. 10, World Scientific, 2019, pp. 169–187.

[10] Neha Bansal, Arun Sharma, R.K. Singh, A review on the application of deep learning in legal domain, in: John MacIntyre, Ilias Maglogiannis, Lazaros Iliadis, Pimenidis Elias (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Switzerland, 2019, pp. 374–381.

[11] Vinod B. Shidham, Vanlila K. Swami, Evaluation of apoptotic leukocytes in peripheral blood smears, Arch. Pathol. Lab Med. 124 (9) (2000) 1291–1294. PMID: 10975923.

[12] Nataša Petrovi¢, Gabriel Moyà-Alcover, Antoni Jaume i Capó, Manuel González-Hidalgo, Sickle-cell disease diagnosis support selecting the most appropriate machine learning method: towards a general and interpretable approach for cell morphology analysis from microscopy images, Comput. Biol. Med. 126 (2020) 104027.

[13] Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard Maude, Stefan Jaeger, Sameer Antani, Deep learning for smartphone-based malaria parasite detection in thick blood smears, IEEE Journal of Biomedical and Health Informatics 99 (1) (2019).

[14] Charles B. Delahunt, Mayoore S. Jaiswal, Matthew P. Horning, Samantha Janko, Clay M. Thompson, Sourabh Kulhare, Liming Hu, Travis Ostbye, Grace Yun, Roman Gebrehiwot, Benjamin K. Wilson, Earl Long, Stephane Proux, Dionicia Gamboa, Peter Chiodini, Jane Carter, Mehul Dhorda, David Isaboke, Bernhards Ogutu, Wellington Oyibo, Elizabeth Villasis, Kyaw Myo Tun, Christine Bachman, David Bell, Courosh Mehanian, Fully-automated Patient-Level Malaria Assessment on Field-Prepared Thin Blood Film Microscopy Images, Including Supplementary Information. *CoRR*, abs/1908.01901, 2019.

[15] Vijayalakshmi Arunagiri, B. Rajesh, Deep learning approach to detect malaria from microscopic images, Multimed. Tool. Appl. 79 (2019) 15297–15317.

[16] Singla Neeru, Vishal Srivastava, Deep Learning Enabled Multi-Wavelength Spatial Coherence Microscope for the Classification of Malaria-Infected Stages with Limited Labelled Data Size. *CoRR*, abs/1903.06056, 2019.

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

[18] Sarmad Shafique, Samabia Tehsin, Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks, Technol. Canc. Res. Treat. 17 (2018) 1–7.

[19] Luis H.S. Vogado, Rodrigo M.S. Veras, Flavio H.D. Araujo, R.V. Silva Romuere, R.T. Aires Kelson, Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification, Eng. Appl. Artif. Intell. 72 (C) (2018) 415–422.

[20] Ying Liu, Feixiao Long, Acute Lymphoblastic Leukemia Cells Image Analysis with Deep Bagging Ensemble Learning, bioRxiv, 2019.

[21] Muhammad Shahzad, Arif Umar, Muazzam Khan, Shirazi Syed, Zakir Khan, Waqas Yousaf, Robust method for semantic segmentation of whole-slide blood cell microscopic images, Computational and Mathematical Methods in Medicine (2020), 1–13, 01 2020.

[22] N. Christy Evangeline, M. Annalatha, Computer aided system for human blood cell identification, classification and counting, in: 2018 Fourth International Conference on Biosignals, Images and Instrumentation (ICBSII), 2018, pp. 206–212.

[23] Dheeraj Mundhra, Bharath Cheluvaraju, Jaiprasad Rampure, Tathagato" Rai Dastidar, Analyzing microscopic images of peripheral blood smear using deep learning, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Switzerland, 2017, pp. 178–185.

[24] J.L. Wang, A.Y. Li, M. Huang, A.K. Ibrahim, H. Zhuang, A.M. Ali, Classification of white blood cells with patternnet-fused ensemble of convolutional neural networks (pecnn), in: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2018, pp. 325–330.

[25] Qiwei Wang, Shusheng Bi, Minglei Sun, Yuliang Wang, Di Wang, Shaobao Yang, Deep learning approach to peripheral leukocyte recognition, PloS One 14 (6) (2019).

[26] Feiwei Qin, Nannan Gao, Yong Peng, Zizhao Wu, Shuying Shen, Artur Grudtsin, Fine-grained leukocyte classification with deep residual learning for microscopic images, Comput. Methods Progr. Biomed. 162 (8) (2018) 243–252.

[27] Xin Geng, Kate Smith-Miles, Incremental Learning, Springer US, Boston, MA, 2009, pp. 731–735.

[28] Deboleena Roy, Priyadarshini Panda, Kaushik Roy, Tree-cnn: A Deep Convolutional Neural Network for Lifelong Learning. *ArXiv*, abs/1802.05800, 2018.

[29] Syed Sarwar, Aayush Ankit, Kaushik Roy, Incremental Learning in Deep Convolutional Neural Networks Using Partial Network Sharing, IEEE Access, 2017, p. 12.

[30] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás guil, cordelia schmid, and karteek alahari. End-to-end incremental learning, in: Yair Weiss (Ed.), Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and, Computer Vision – ECCV 2018, Springer International Publishing, Switzerland, 2018, pp. 241–257.

[31] Roxana Istrate, A. Cristiano I. Malossi, Costas Bekas, Dimitrios Nikolopoulos, Incremental training of deep convolutional neural networks, in: Proceedings of the International Workshop on Automatic Selection, Configuration and Composition of Machine Learning Algorithms, 09 2017, pp. 41–48.

[32] Y. Tao, Y. Tu, M. Shyu, Efficient incremental training for deep convolutional neural networks, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 286–291.

[33] Ademola Adewoyin, Benedict Nwogoh, Peripheral blood film - a review, Ann. Ib. Postgrad. Med. 12 (71–9) (2014) 12.

[34] Mukesh Saraswat, K.V. Arya, Automated microscopic image analysis for leukocytes identification: a survey, Micron 65 (2014) 20–33.

[35] K.K. Anilkumar, V.J. Manoj, T.M. Sagi, A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of leukemia, Biocybernetics and Biomedical Engineering 40 (4) (2020) 1406–1420.

[36] Philip Lanzkowsky, Chapter 11 - disorders of white blood cells, in: Philip Lanzkowsky, Editor, *Manual Of Pediatric Hematology And Oncology*, fifth ed., Academic Press, San Diego, 2011, pp. 272–320, fifth edition edition.

[37] G. Zini, Abnormalities in leukocyte morphology and number, Blood and Bone Marrow Pathology (01 2011) 247–261.

[38] M. Chen, X. Shi, Y. Zhang, D. Wu, M. Guizani, Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network, IEEE Transactions on Big Data, 2017, 1–1.

[39] W. Pan, Yuhang Dong, Dongsheng Wu, Classification of Malaria-Infected Cells Using Deep Convolutional Neural Networks, vol. 8, IntechOpen, 2018, pp. 159–172, 09.

[40] Sivaramakrishnan Rajaraman, Stefan Jaeger, K. Sameer, Antani, Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images, PeerJ 7 (e6977) (2019).

[41] J. Lemley, S. Bazrafkan, P. Corcoran, Smart augmentation learning an optimal data augmentation strategy, IEEE Access 5 (2017) 5858–5869.

[42] R. Al-Qudah, C.Y. Suen, Synthetic blood smears generation using locality sensitive hashing and deep neural networks, IEEE Access 8 (2020) 102530–102539.

[43] Ruggero Labati, Vincenzo Piuri, Fabio Scotti, All-idb: the acute lymphoblastic leukemia image database for image processing, in: 18th IEEE International Conference on Image Processing, Sept 2011, pp. 2045–2048. Brussels, Belgium.

[44] Cecilia Di Ruberto, Andrea Loddo, Lorenzo Putzu, A Region Proposal Approach for Cells Detection and Counting from Microscopic Blood Images, 2019, pp. 47–58, 09.

[45] Cecilia Di Ruberto, Andrea Loddo, and Giovanni Puglisi. Blob detection and deep learning for leukemic blood image analysis. Appl. Sci., 10(3), 2020.

[46] Chenyi Chen, Ming-Yu Liu, Oncel Tuzel, Jianxiong Xiao, R-cnn for small object detection, in: Vincent Lepetit, Nishino Ko, Yoichi Sato (Eds.), Shang-Hong Lai, Springer International Publishing, Switzerland, 2017, pp. 214–230. Computer Vision – ACCV 2016.

[47] Luzio Tremolizzo, Gessica Sala, Carlo Ferrarese, Platelet Activation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 1034–1035.

[48] C.Y. Suen, R. Al-qudah, A Survey on Peripheral Blood Smear Analysis Using Deep Learning, vol. 63, Springer, Switzerland, 2020, pp. 725–738, 11.
[49] Keras, Keras applications. https://keras.io/api/applications/, 2020.
[50] M. Shanker, M.Y. Hu, M.S. Hung, Effect of data standardization on neural network training, Omega 24 (4) (1996) 385–397.
[51] Jesús Angulo and Georges Flandrin. Automated detection of working area of peripheral blood smears using mathematical morphology. Anal. Cell Pathol.: the journal of the European Society for Analytical Cellular Pathology, 25:37–49, 02 2003.
[52] S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, K.E. Muller, Contrast-limited adaptive histogram equalization: speed and effectiveness, in: [1990] Proceedings of the First Conference on Visualization in Biomedical Computing, 1990, pp. 337–345.
[53] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.
[54] M. Khodashenas, H. Ebrahimpour-komleh, A.M. Nickfarjam, White blood cell detection and counting based on genetic algorithm, in: 2019 Advances in Science and Engineering Technology International Conferences (ASET), 2019, pp. 1–4.
[55] Alferez Santiago, Anna Merino, Andrea Acevedo Lipes, Laura Puigvi, Jose Rodellar, Color clustering segmentation framework for image analysis of malignant lymphoid cells in peripheral blood, Med. Biol. Eng. Comput. 57 (2019) 2.
[56] Dheeraj Mundhra, Bharath Cheluvaraju, Jaiprasad Rampure, Tathagato Rai Dastidar, Analyzing microscopic images of peripheral blood smear using deep learning, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Switzerland, 2017, pp. 178–185.
[57] Salim Arslan, Emel Ozyurek, Cigdem Gunduz-Demir, A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images, Cytometry 85 (6) (2014) 480–490.
[58] Xueyin Lin, W. Wee, Shape detection using range data, Proceedings. 1985 IEEE International Conference on Robotics and Automation 2 (1985) 34–39.
[59] University of Leeds, White blood cells. https://www.histology.leeds.ac.uk/blood/bloodwbc.php, 2020.
[60] EclinPath, Normal leukocytes. https://eclinpath.com/hematology.morphologic-features/white-blood-cells/normal-leukocytes/, 2020.
[61] Talaat Ahmed, Philip Kollmannsberger, Ewees Ahmed, Efficient classification of white blood cell leukemia with improved swarm optimization of deep features, Sci. Rep. 10 (2020) 2.
[62] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.
[63] Yahui Jiang, Meng Yang, Shuhao Wang, Xiangchun Li, Yan Sun, Emerging role of deep learningbased artificial intelligence in tumor pathology, Canc. Commun. 40 (2020) 154–166.
[64] Masoumeh Zareapoor, Pourya Shamsolmoali, Jain Deepak Kumar, Haoxiang Wang, Jie Yang, Kernelized support vector machine with deep learning: an efficient approach for extreme multiclass dataset, Pattern Recogn. Lett. 115 (2018) 4–13. Multimodal Fusion for Pattern Recognition.
[65] Gonzalo D. Sad, Lucas D. Terissi, Juan C. Gómez, Class confusability reduction in audio-visual speech recognition using random forests, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications,, Springer International Publishing, Switzerland, 2018, pp. 584–592.
[66] Maya R. Gupta, Samy Bengio, Jason Weston, Training highly multiclass classifiers, J. Mach. Learn. Res. 15 (43) (2014) 1461–1492.
[67] Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, Zheng Zhang, Error-driven incremental learning in deep convolutional neural network for large-scale image classification, in: ACM Multimedia, November 2014, pp. 177–186.
[68] Aditya Devarakonda, Maxim Naumov, Michael Garland, Adabatch: Adaptive Batch Sizes for Training Deep Neural Networks, 2018.
[69] Keras, Keras faq. https://keras.io/gettingstarted/faq/#how-can-i-obtain-reproducible-results-using-keras-during-development, 2020.
[70] Riach Duncan, Determinism in deep learning. https://developer.nvidia.com/gtc/2019/video/S9911, 2019.
[71] Yu Zhang, Cangzhi Jia, Keong Kwoh Chee, Predicting the interaction biomolecule types for lncRNA: an ensemble deep learning approach, Briefings Bioinf. 10 (2020).
[72] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Reed Scott, Cheng-Yang Fu, C. Alexander, Berg Ssd, Single shot multibox detector, in: Computer Vision – ECCV 2016, Springer International Publishing, Switzerland, 2016, pp. 21–37.
[73] Rabiah Al-qudah, Ching Y. Suen, Enhancing yolo deep networks for the detection of license plates in complex scenes, in: Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems, DATA '19, New York, NY, USA, Association for Computing Machinery, 2019.

Rabiah A. Al-Qudah Is currently a PhD candidate in computer science at Concordia university. She earned her Msc degree in computer science in 2018 from Concordia university, and her BSc degree in computer science from Jordan University of Science and Technology.She has seven years of experience in software and database development. She published several articles in international conferences and journals. Her research interest includes deep learning and computer aided diagnosis. She was awarded the best paper award in ICVISP 2019She has seven years of experience in software and database development. She published several articles in international conferences and journals. Her research interest includes deep learning and computer aided diagnosis. She was awarded the best paper award in ICVISP 2019.

Ching Y. Suen received the M.S. degree in electronics from the University of Hong Kong, Hong Kong, and the Ph.D. degree in man–computer communications from the University of British Columbia, Vancouver, BC, Canada. After graduation, he joined the Department of Computer Science, Concordia University, Montreal, QC, Canada, where he served as the Chairman, and the Associate Dean (Research) with the Faculty of Engineering and Computer Science. He is currently the Director of the Center for Pattern Recognition and Machine Intelligence, Concordia University. He has supervised 120 doctoral and master's students to completion, and guided/hosted 100 long-term visiting scientists and professors. He has always been fascinated by letters and characters, ever since he started his doctoral research on teaching the computer to read multifont documents with a voice output for the blind. He has authored or co-authored 6 conference proceedings, 15 books, and more than 550 papers. Dr. Suen became a Fellow of the IEEE in 1986, a Fellow of the IAPR in 1994 and the Academy of Sciences of the Royal Society of Canada in 1995. He has served at numerous national and international professional societies as the President, the Vice President, the Governor, and the Director. He is the Founder of four conferences, such as the International Conference on Document Analysis and Recognition (ICDAR), International Workshop/Conference on Frontiers in Handwriting Recognition (ICFHR), Vision Interface (ICPRAI), and has also organized numerous international conferences, including International Conference on Pattern Recognition, ICPR, ICDAR, ICFHR, ICPRAI and International Conference on the Computer Processing of Oriental Languages. He was a recipient of numerous awards, including the Gold Medal from the University of Bari, Italy, in 2012, the IAPR ICDAR Award in 2005, the ITAC/NSERC National Award in 1992, and the Concordia Fellow Award and the Concordia Lifetime Research Achievement Award in 1998 and 2008. In 1997, he created the IAPR ICDAR Awards, to honor both young and established outstanding researchers in the field of document analysis and recognition. He had been the Editor-in-Chief of Pattern Recognition for 10 years and became the Emeritus EIC in 2018. Presently, he is an Associate Editor for several journals.Dr. Suen was awarded the King-Sun Fu prize in January 2021.