Received: 29 November 2023 | Revised: 14 March 2024 | Accepted: 25 March 2024

DOI: 10.1002/cyto.a.24839

ORIGINAL ARTICLE

ISAC CYTOMETRY Journal of Quantitative Cell Science PART A

Comprehensive data analysis of white blood cells with classification and segmentation by using deep learning approaches

Şeyma Nur Özcan[1] | Tansel Uyar[1] | Gökay Karayeğen[2]

[1]Biomedical Engineering Department, Başkent University, Ankara, Turkey

[2]Biomedical Equipment Technology, Vocational School of Technical Sciences, Başkent University, Ankara, Turkey

Correspondence
Gökay Karayeğen, Biomedical Equipment Technology, Vocational School of Technical Sciences, Başkent University, Bağlıca Campus, 06790, Ankara, Turkey.
Email: karayegengokay@gmail.com

Abstract

Deep learning approaches have frequently been used in the classification and segmentation of human peripheral blood cells. The common feature of previous studies was that they used more than one dataset, but used them separately. No study has been found that combines more than two datasets to use together. In classification, five types of white blood cells were identified by using a mixture of four different datasets. In segmentation, four types of white blood cells were determined, and three different neural networks, including CNN (Convolutional Neural Network), UNet and SegNet, were applied. The classification results of the presented study were compared with those of related studies. The balanced accuracy was 98.03%, and the test accuracy of the train-independent dataset was determined to be 97.27%. For segmentation, accuracy rates of 98.9% for train-dependent dataset and 92.82% for train-independent dataset for the proposed CNN were obtained in both nucleus and cytoplasm detection. In the presented study, the proposed method showed that it could detect white blood cells from a train-independent dataset with high accuracy. Additionally, it is promising as a diagnostic tool that can be used in the clinical field, with successful results in classification and segmentation.

KEYWORDS

CNN, image classification, independent dataset, nucleus and cytoplasm segmentation, white blood cells

1 | INTRODUCTION

Human blood consists of various cells that perform functions such as oxygen transport, coagulation, and protection/defense of the body against various viruses and microbes, and immunity. White blood cells (WBCs) protect the body against various microbes, viruses, infections, and pathogens. An adult human has 4000 to 10,000 WBCs per microliter (μL) of blood. It is possible to make various inferences when WBCs are evaluated both quantitatively and qualitatively [1]. In quantitative evaluation, the fact that WBC is outside these values is used as an indicator of various diseases, such as leukemia, infectious diseases, liver diseases, and bone marrow failure. On the other hand, various information can be obtained from the qualitative (morphological) evaluations of WBCs. An underlying infection may be the result of any aberrant WBC morphology, or premature leukocytes are another example of abnormal morphology of the WBC. These are called blast cells and are seen in conditions such as leukemia. Detection of these conditions in WBC morphologies requires the determination of

parameters such as nucleus-cell ratio, number of nuclei, nucleus size, and so forth [2, 3]. For this reason, monitoring the health of our defense system is of critical importance for the detection and early intervention of the conditions mentioned above. Therefore, the classification and segmentation of WBCs in the blood is a significant task.

WBCs consist of three parts: cytoplasm, cell wall, and nucleus. There are five types of WBCs: basophils, eosinophils, lymphocytes, monocytes, and neutrophils. These five types fulfill the task of protecting and defending the body against viruses, bacteria, and infections. Each subsection has a certain limit value or morphology [1]. Passing these limit values or having abnormal morphology can indicate the presence of diseases. For this reason, WBCs must be classified and segmented quickly and accurately. In classical methods, laboratory attendants used a microscope to determine the classes of WBCs by looking at the stained cells. This method requires a lab expert to have much clinical experience and causes various errors. Today, these errors can be minimized with methods that include both image processing and artificial intelligence approaches. In addition, rapid results can be obtained in the completion of these analyses, which have a high workload.

Today, the current and common of these methods are approaches that include deep learning. In the literature, there are studies on segmentation as well as studies on classification, but there are also studies that include both. In these studies, in the literature, different data sources were used with very different methodological approaches.

One of the most common methods of studies is to evaluate the performance of innovative methods using a single dataset. In this group of studies, the known models in the literature were tested first [4–11]. For example, in a study conducted by Macawile et al. [12], AlexNet, GoogLeNet, and ResNet-101, which are pretrained networks from convolutional neural networks (CNNs), were used. These pretrained networks were tested in the ALL-IDB database, and the best result was obtained in AlexNet with 96.63% accuracy. In another study conducted by Bagido et al. [13], WBCs were classified by using pretrained networks and various optimization methods together. Pretrained networks VGG-16, Xception, MobileNetV2, InceptionResNetV2, and optimization methods Stochastic Gradient Descent (SGD) and Adam (Adaptive Moment Estimation) were used. As a result of the studies, it was seen that the best result was in the combination of InceptionResNetV2 and SGD with 96.21% accuracy. In a different study by Khouani et al. [14] for WBC segmentation, they used ResNet-50, ResNet-101, VGG-16, VGG-19, and Inception-v3 networks. As a result of the study, the best test accuracy was 95.73% in segmentation for ResNet50 with the Adam optimizer. In single dataset studies, methods such as hybridizing the models or combining them with different classifier models for feature extraction as another group of approaches are used [15–23]. In the study conducted by Toğaçar et al. [24], AlexNet, VGG-16, and LeNet, which are pretrained networks, were applied to classify WBCs. Along with these pretrained networks, various classifiers, such as decision tree, support vector machine linear discriminant analysis, quadratic discriminant analysis, nearest neighbor, and softmax, were used as a hybrid model. The best results were obtained in the hybrid model of AlexNet and quadratic discriminant analysis, with an accuracy of 97.78%. In addition to all these approaches, some studies build their own architectures [25–27].

Contrary to the studies discussed above, there are also studies that have applied various approaches to use more than one dataset separately [28–33]. In the study conducted by Çınar et al. [33] in 2021, AlexNet-GoogLeNet-SVM (support vector machine) methods were used as a hybrid for the classification of WBCs. The created network was tested on two different datasets, Kaggle and LISC. The results obtained were created for two separate datasets. The accuracy obtained on the Kaggle dataset is 99.73%, and the accuracy obtained on the LISC dataset is 98.23%. Only one study in the literature has combined two datasets [34]. Details of the studies summarized above are presented in the tables below (Tables 1 and 2).

The common feature of related studies, which include classification, segmentation, or both, was that they used more than one dataset, but used them separately [28–33]. No study has been found in the literature that contains more than two datasets that are combined and used together. In the present study, a mixed dataset was obtained by combining four separate datasets. In the first step of the present study, a classification model was produced using a deep learning method, CNN, to specify WBC types in given blood samples. Using this produced model and the mixed dataset, training and testing processes were run for five classes (basophils, eosinophils, lymphocytes, monocytes, and neutrophils) in classification. The purpose of using a mixed dataset in the study was to produce a robust classifier that can parse data from any source. As the second step of the study, a completely independent dataset, which was not used in the training phase, was created to test the mentioned robust classification ability. This train-independent dataset consists of the combination of two separate datasets, and a second test phase was performed with the produced model. This approach is believed to be very useful for future applications. The other aim of the study is to develop a segmentation model that is able to produce outputs that can be used in qualitative analysis, such as morphological assessment. Consequently, the segmentation of WBCs is significant in analyzing the interaction between cytoplasm and nucleus for numerous reasons, particularly in the context of understanding the shape and functionality of these immune cells. The identification of morphological anomalies in the cytoplasm and nucleus is made possible by accurate segmentation. This is important because variations in the size, shape, or lobulation of the nucleus and in cytoplasmic characteristics can signal particular health issues including infections, leukemia, and other blood disorders. Each WBC type has different morphological traits, such as shape, texture, size, and form of cytoplasm and nucleus, which could manifest as one or more lobes based on the interaction of their unique granules with a staining process. The counting, segmentation, and classification of WBCs can be facilitated by these properties [35, 36]. With this point of view, in addition to classification, a morphological examination tool that does not contribute to classification but only facilitates visual assessment was presented. A semantic segmentation model with CNN was produced to determine WBC nuclei and cytoplasm parts. At the same time, two different pretrained models, SegNet and UNet, were used to perform a more comprehensive assessment and to obtain a detailed analysis. The results of these methods were compared to the proposed CNN segmentation method.

TABLE 1 Brief summary of white blood cells (WBC) single-dataset studies for classification and segmentation.

Reference	Single dataset	Method	Accuracy rate
Hegde et al. [18]	Private dataset	Classification (4 Class) with Proposed Hybrid Method	99.5%
Khouani et al. [14]	CHU Tlemcen (Private Dataset)	Segmentation with Pre-trained Models	ResNet-50: 95.73%
Aziz et al. [5]	ALL-IDB	Classification (5 Class) with Pre-trained Models	AlexNet: 91.93% ResNet-18: 93.3%
Macawile et al. [12]	ALL-IDB	Classification (5 Class) with Pre-trained Models	AlexNet: 96.63% GoogLeNet: 97.54% ResNet-101: 97.53%
Shahin et al. [25]	ALL-IDB	Classification (5 Class) with Proposed Method	96.1%
Ridoy et al. [4]	Kaggle	Classification (4 Class) with Pre-trained Model	Modified LeNet: 95.97%
Siddique et al. [9]	Kaggle	Classification (4 Class) with Pre-trained Model	SqueezeNet: 93.80%
Toğaçar et al. [15]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	97.95%
Yildirim et al. [19]	Kaggle	Classification (4 Class) with Pre-Trained Models	AlexNet: 80.77% ResNet-50: 80.02% DenseNet-201: 83.44% GoogLeNet: 75.21%
Mohamed et al. [20]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	92%
Patil et al. [21]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	95.89%
Khan et al. [22]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	99.12%
Praveen et al. [23]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	90%
Toğaçar et al. [24]	Kaggle	Classification (4 Class) with Proposed Hybrid Method	97.78%
Girdhar et al. [26]	Kaggle	Classification (4 Class) with Proposed Method	98.55%
Ma et al. [8]	DCGAN-generated with Matrix transformation (Source: Kaggle)	Classification (4 Class) with Pre-trained Model	ResNet + Transfer Learning + Improved Loss Function: 91.68%
Merino et al. [10]	CellaVision	Classification (Normal/Abnormal Leukocytes) with Pre-trained Model	VGG-16: 96%
Shaheen et al. [11]	CellaVision	Classification (Normal/Abnormal Monocytes and Leukocytes) with Pre-trained Models	LeNet-5: 96.32% AlexNet: 98.58%
Bagido et al. [13]	CellaVision	Classification (5 Class) with Pre-trained Models	MobileNetV2: 79.79% VGG-16: 90.28% Xception: 90.37% InceptionResNetV2: 96.21%
Gavas et al. [17]	CellaVision	Classification (8 Class) with Proposed Hybrid Method	99.51
Mohammadi et al. [27]	CellaVision	Segmentation with Proposed Method	Nucleus: 95% Cytoplasm: 91%
Hartanto et al. [6]	DCGAN-generated (Source: CellaVision)	Classification (5 Class) with Pre-trained Model	ResNet-50: 82.5%
Almezhghwi et al. [7]	LISC	Classification (5 Class) with Pre-trained Model	VGG-16: 90.6% VGG-19: 91.8% ResNet-18: 91.1% ResNet-50: 92.7% DenseNet-121: 93.9% DenseNet-169: 94.4%
Reena et al. [16]	LISC	Classification (5 Class) with Pre-trained Models	AlexNet: 98.87%

Each trained segmentation network (proposed CNN, SegNet, UNet) was tested on the mixed datasets that was produced prior. All trained networks were evaluated according to the performance evaluation metrics determined for classification and segmentation.

2 | MATERIALS AND METHODS

In this section, information about the dataset used and the proposed methods are detailed.

TABLE 2 Brief summary of white blood cells (WBC) multi-dataset studies for classification and segmentation.

Reference	Multi dataset	Method	Accuracy rate
Long et al. [28]	CellaVision, Kaggle, ALL-IDB2	Classification (Various Classes) with Proposed Method (BloodCaps)	CellaVision: 97.29% Kaggle: 99.47% All-IDB2: 99.34%
Yao et al. [29]	Private Dataset, Kaggle	Classification (4 Class) for Noisy or Low Resolution Datasets with Proposed Method (TWO-DCNN)	Private Dataset: 95.1% Kaggle: 91.6%
Roy et al. [30]	JTSC, CellaVision, LISC	Segmentation with Proposed Method (DeepLabv3+)	JTSC: 99.15% CellaVision: 97.40% LISC: 98.22%
Lu et al. [31]	JTSC, CellaVision, LISC, Private Dataset	Segmentation with Proposed Method (WBC-Net)	JTSC: 98.21% CellaVision: 98.33% LISC: 94.20% Private Dataset: 94.35%
Banik et al. [32]	Kaggle, ALL-IDB2, JTSC, CellaVision	Segmentation (Nucleus) with Proposed Method	Kaggle: 99.42% ALL-IDB2: 98.61% JTSC: 97.57% CellaVision: 98.86%
Çınar et al. [33]	Kaggle, LISC	Classification (4 Class) with AlexNet-GoogLeNet-SVM Concatination Model	Kaggle: 99.73% LISC: 98.23%
Kutlu et al. [34]	Kaggle and LISC	Classification (5 Class) with R-CNN Method of Various Pre-trained Models	97.52%

2.1 | Details of datasets

In the present study, two separate CNN models were developed for both classification and segmentation using the combined form of four separate WBC datasets. Using this combined (mixed) dataset, both the training and the test process were conducted for both classification (for five classes) and segmentation (for four classes). After the test processes, performance evaluations were made. To evaluate the robustness of the produced classification model, completely independent data that were not used in the training process were used. For this independent dataset where robustness was tested, two more datasets were combined and used. This evaluation was both applied for the classification model and segmentation model. Detailed information about the datasets used in the study is given below.

2.1.1 | The datasets used in the training (classification and segmentation) and testing processes

a. CellaVision dataset (Dataset 1)

In its original form, the dataset included 17,092 pictures of distinct healthy cells that were collected using the CellaVision DM96 analyzer in the Core Laboratory of the Hospital Clinic of Barcelona. The blood smears were automatically stained using May Grünwald-Giemsa in the autostainer Sysmex SP1000i. There are eight different types of cells in the original dataset (neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and platelets). The photos are 360 × 363 pixels in size and were annotated by skilled clinical pathologists. The photos were taken from healthy persons who were free of infection, hematologic or oncologic illness at the time of blood collection and receiving no pharmacological treatment [37] (More information can be found at https://data.mendeley.com/datasets/snkd93bnjr/1 Accessed date: 13/02/2023).

b. First and second microscope images of the Raabin-WBC dataset (Dataset 2 & 3)

All samples were collected from the laboratory of Razi Hospital in Rasht, Gholhak Laboratory, Shahr-e-Qods Laboratory and Takht-e Tavous Laboratory in Tehran, Iran. Then, two microscopes were used to image the slides. First, microscopy images were taken by an Olympus CX18 microscope and a Samsung Galaxy S5 camera. Second microscopy images were taken by a Zeiss microscope and an LG G3 camera. Therefore, this dataset contains images obtained from two separate sources (datasets). A total of 72 healthy peripheral blood films (male and female samples from ages 12 to 70) were used to collect images of neutrophils, eosinophils, monocytes, and lymphocytes. The blood smears were stained using May Grünwald-Giemsa in both datasets. Peripheral blood cell films were tagged and filed by two expert analysts. The tagged images were cropped from the large image by the researchers according to the tag names and saved in the necessary folders [38]. The original files contained data from a total of 17,965 WBCs (More information can be found at https://raabindata.com/free-data/ Accessed date: 13/02/2023).

c. Kaggle blood cell images dataset (Dataset 4)

The original dataset contained 12,500 augmented images of blood cells (JPEG) with accompanying cell type labels. The Gismo-right

technique has been employed for staining the blood smears. The augmented dataset was not used directly. This dataset was accompanied by a second dataset that contained the original 410 pictures (preaugmentation). In more detail, the "dataset-master" folder contained 410 photos of blood cells with bounding boxes and subtype labels. Eosinophils, lymphocytes, monocytes, neutrophils and basophils are the different cell types in this dataset [39] (More information can be found at https://www.kaggle.com/datasets/paultimothymooney/blood-cells Accessed date: 13/02/2023).

2.1.2 | The datasets used for evaluating the robustness of the produced classification model

a. LISC dataset (Dataset 5)

The hematological pictures obtained from healthy patients' peripheral blood were stored in the LISC database. Eight healthy volunteers' peripheral blood was sampled, and 400 samples were collected from 100 microscope slides. Gismo-Right was used to smear and stain the microscope slides. Images from the stained peripheral blood were then captured using a light microscope (Microscope-Axioskope 40) and an achromatic lens with 100× magnification. Then, using a digital camera with the model number SSCDC50AP from Sony, these pictures were taken and saved in BMP format. The images contain 720 × 576 pixels. They are all color photographs that were gathered from Tehran, Iran's Imam Khomeini Hospital's Hematology-Oncology and BMT Research Center. The images were classified by a hematologist into normal leukocytes: basophils, eosinophils, lymphocytes, monocytes, and neutrophils. The sections with WBC cells were cropped from the large slide image and recorded as smaller pieces [40] (More information can be found at http://users.cecs.anu.edu.au/~hrezatofighi/Data/Leukocyte%20Data.htm Accessed date: 13/02/2023).

b. Blood cell detection (BCD) dataset (Dataset 6)

In this dataset, WBCs and red blood cells (RBCs) from peripheral blood smears photographed under a light microscope are annotated. Romanowsky staining was used to smear and stain the microscope slides. Contains 100 annotated images in png format, with 2237 labeled RBCs and only 103 labeled WBCs. Each image consists of 256 pixels in height and width of RGB channels. Annotations.csv file contains both locations and labels [41] (More information can be found at https://github.com/draaslan/blood-cell-detection-dataset Accessed date: 13/02/2023).

2.1.3 | Examining and combining datasets

Before the datasets were combined, they were subjected to an elimination process. Images of cells with exploded cytoplasm and deformed form and images with more than one cell in the same frame were excluded from the combined dataset. After the elimination process, the final mixed dataset (CellaVision, First and Second Raabin-WBC, Kaggle) was obtained. A total of 14,563 photos of five categories of WBC cells (neutrophils, eosinophils, basophils, lymphocytes, and monocytes) were used in the training and test processes of the classification study. In the segmentation study, four types of WBC cells (neutrophils, eosinophils, lymphocytes, and monocytes) were selected from the mixed dataset for nuclei and cytoplasm segmentation, which contains 13,105 images. Basophil cells were not included in the segmentation studies because the morphology of the nucleus covered the cytoplasm; in other words, the separation of the cytoplasm and the nucleus was unclear. A total of 323 photos (combined LISC and BCD) of five categories of WBC cells (neutrophils, eosinophils, basophils, lymphocytes, and monocytes) were used to evaluate the robustness of the produced classification model.

All images of WBCs were resized to 360 × 360 pixels in png format. The images were not filtered or preprocessed in any way. Each dataset is split into 56% for training, 14% for validation and 30% for testing. Then, all training, validation and testing folders were combined. Labeling was performed using the Ground Truth Labeler tool in MATLAB for segmentation. In this way, ground truth (nuclei, cytoplasm, and background sublabels) images were obtained. The numerical distribution information of all datasets used in the study is detailed in Table 3.

On the other hand, the datasets were used individually due to understand the performance of the classification model presented in the study and to compare it with the performance of other models in the literature. For this reason, separate training and testing processes were evaluated for Dataset 1, Dataset 2 & 3 and Dataset 4. The above-mentioned ratios were used for training, validation and testing. In order to better understand the generalization performance of the combined data set, which constitutes the main motivation of the study, the train-independent dataset was tested with trained models obtained by using separate data sets. Dataset 1, Dataset 2 & 3 and Dataset 4 were used separately during the training process, and no data was allocated from these datasets for testing.

Data augmentation was used to increase training accuracy and prevent overfitting in classification studies. By using the augmentation function, the size of the dataset was increased with the operation of rotation (range: ±20°), reflection (in both the x-axis and y-axis), and translations (in both the x-axis and y-axis with distance: ±3) within the training process. Therefore, the "augmentedImageDatastore" structure was used in MATLAB. The data augmentation method defined in the MATLAB platform operates with a different method compared to the traditional usage known in the literature. In traditional methods, after data augmentation, the data in the datasets are increased physically, and the changed data are saved in the dataset as new data. On the other hand, in the MATLAB platform, new data produced by using augmented parameters are used to feed the neural network through mini-batches in each epoch without being physically saved. Thus, the augmented image datastore modifies images in the mini-batch of training data in a random manner for each training cycle. The training

TABLE 3 Distribution of the white blood cells (WBC) datasets.

	Training and testing dataset				Training-independent dataset
Data	Training (56%)	Validation (14%)	Testing (30%)	Total (100%)	Testing robustness of classification and segmentation model (100%)
Basophil	816	204	438	**1458**	53
Eosinophil	2306	576	1220	**4102**	72
Lymphocyte	3199	800	1713	**5712**	71
Monocyte	1284	321	688	**2293**	52
Neutrophil	8204	2051	4308	**14,563**	75
Total	**15,809**	**3952**	**8367**	**28,128**	**323**

data for each epoch are randomly shuffled when an augmented image datastore is used as the source of training images. Therefore, a slightly different dataset was created for each epoch. The modified dataset technically alters existing images, but it will provide distinct data for the CNN model, which prevents the CNN model from overfitting and memorizing the exact information of the training images. In conclusion, apart from the data augmentation process used in the training process and resizing process for datasets, there is no other preprocessing method in the test or training process.

2.2 | Convolutional neural networks

In our study, the dataset detailed above was used for both classification and segmentation. Two separate structures are designed for both classification and segmentation. The schematic detailing of the presented study is given in Figure 1 below.

In addition, the details of these designed architectures are given in the following sections. For both experimental studies, the same computer setup was used, which has 2 Intel Xeon E5620 processors, an NVIDIA GTX 1060 6 GB graphics card with a capability of 4.4 TFLOPS, and 16 GB of RAM with 1066 MHz. Additionally, our experiments were implemented in MATLAB 2022b.

2.2.1 | Classification

Using CNNs, which are a type of artificial neural network, images of WBCs were classified. Neutrophils, eosinophils, basophils, lymphocytes, and monocytes were described as five classes. Data augmentation was implemented on the resized dataset at the very beginning. The data were transformed into a new dataset for each epoch using the random augmentation parameters described above. After that, the altered dataset covering all these classes was fed into the neural network as an input. In addition, the convolution layer (filter size of 5×5) (Conv) followed by the batch normalization layer (BN), leaky ReLU (leaky rectified linear unit) layer, and average pooling layer (stride step with two and pool size of 2×2) was repeated seven times in the proposed method. A fully connected layer (FC) and a softmax layer were then added in the next phase. From the characteristics that had been collected by all the preceding layers, the classification layer was utilized to categorize the type of WBC cell for the finalization. The hierarchical structure of the proposed method is detailed below in Figure 2. The proposed model's complexity in terms of FLOPs was 736 MFLOPs with 17.5 M parameters. The inference time of the proposed model was 1.67×10^{-4} s.

In the proposed method, the initial learning rate was set to 0.0003, and the stochastic gradient descent with momentum (SGDM) optimizer was selected for training the network solver. The maximum number of epochs was set to 500 with the every-epoch shuffling option. At the end of the training process, the trained CNN model was validated by validation accuracy. The optimization curve of the trained model with objective functions (loss curve function and accuracy curve function) is shown in Figure 3.

The trained CNN model was tested on test data, and the results were characterized with evaluation metrics (accuracy, etc.) detailed below.

2.2.2 | Segmentation

Unlike in classification, four cell types (neutrophils, eosinophils, lymphocytes, and monocytes) were used to detect nuclei, cytoplasm, and background in segmentation. Data augmentation was not used in this experimental study. Four class images of the dataset and their ground-truth labels were fed into the neural network as an input. In addition to the proposed CNN model, segmentation was performed with the pretrained CNN models UNet and SegNet using the same dataset. Then, performance metrics were compared, and comprehensive statistics, including class discrimination and accuracy, were determined.

a. Proposed CNN

The training process of this CNN model consists of convolution layers, batch normalization layers, ReLU layers, and max-pooling layers that were responsible for downsampling as the first part. As the second part, transposed convolution layers and ReLU layers were used for upsampling. At the end of the proposed CNN model, a softmax layer and pixel classification layer were applied. With this step, the

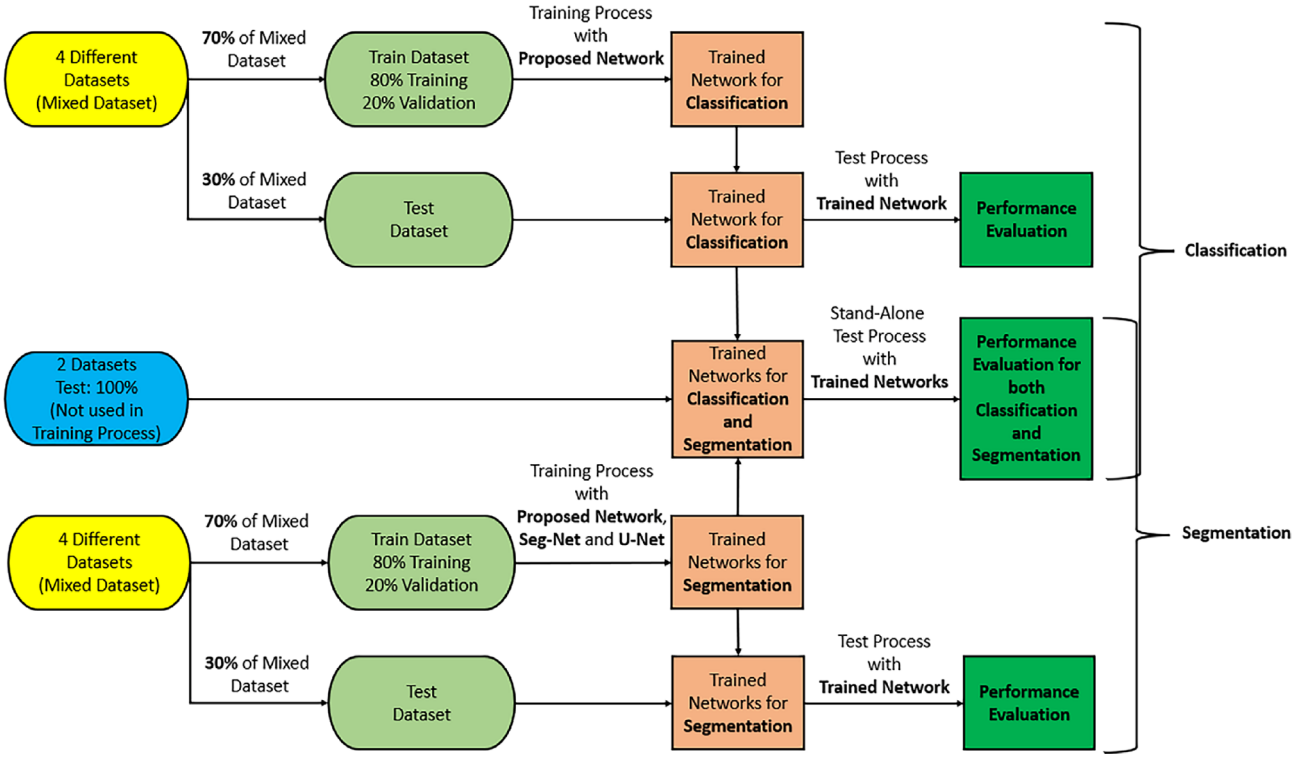

FIGURE 1 Detailed demonstration of presented study. [Color figure can be viewed at wileyonlinelibrary.com]

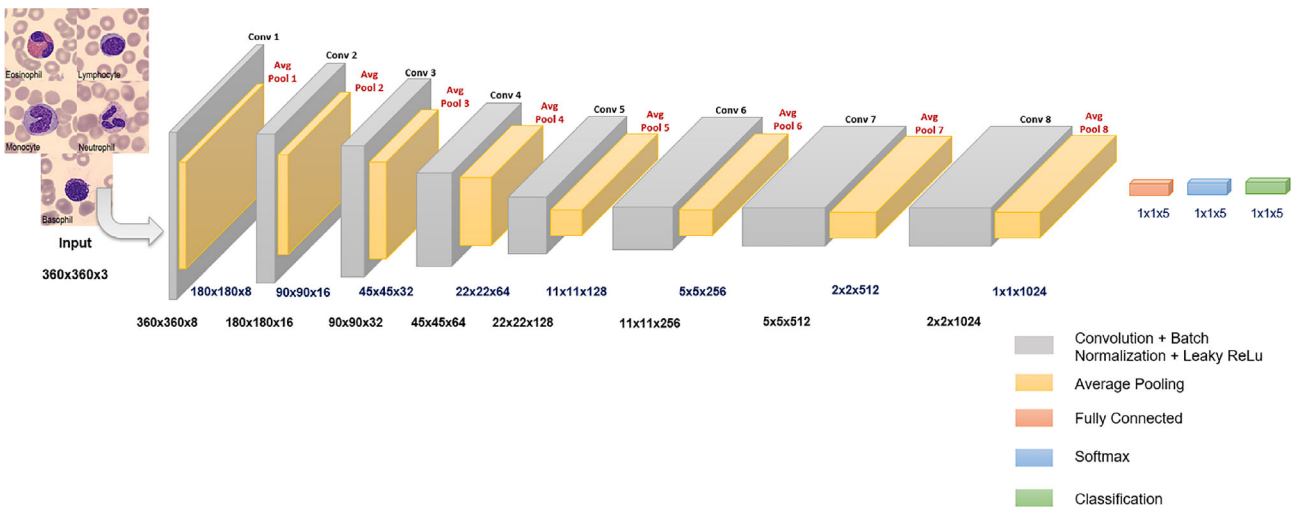

FIGURE 2 The proposed deep learning network for white blood cells (WBC) classification. [Color figure can be viewed at wileyonlinelibrary.com]

training process was completed. Semantic segmentation was used to test images following the training phase. In semantic segmentation, the nucleus, cytoplasm, and background pixels were predicted by the trained CNN model. In the continuation, the results were characterized with evaluation metrics (global accuracy, etc.) detailed below. The hierarchical structure of the proposed method is detailed below in Figure 4. The proposed model's complexity in terms of FLOPs was 7.311 BFLOPs with 58.5 K parameters. The inference time of the proposed model was 1.63×10^{-3} s.

In the designed architecture, the initial learning rate was set to 0.0003, and the SGDM optimizer was selected for training the network solver. Even though SGDM is slower than ADAM in the training phase, it is more efficient and provides higher performance on test data. In this study, we aimed to obtain effective and faster performance in the test phase [42–44].

The maximum number of epochs was set to 10 with every-epoch shuffling option. At the end of the training process, the trained CNN model was validated with validation accuracy.

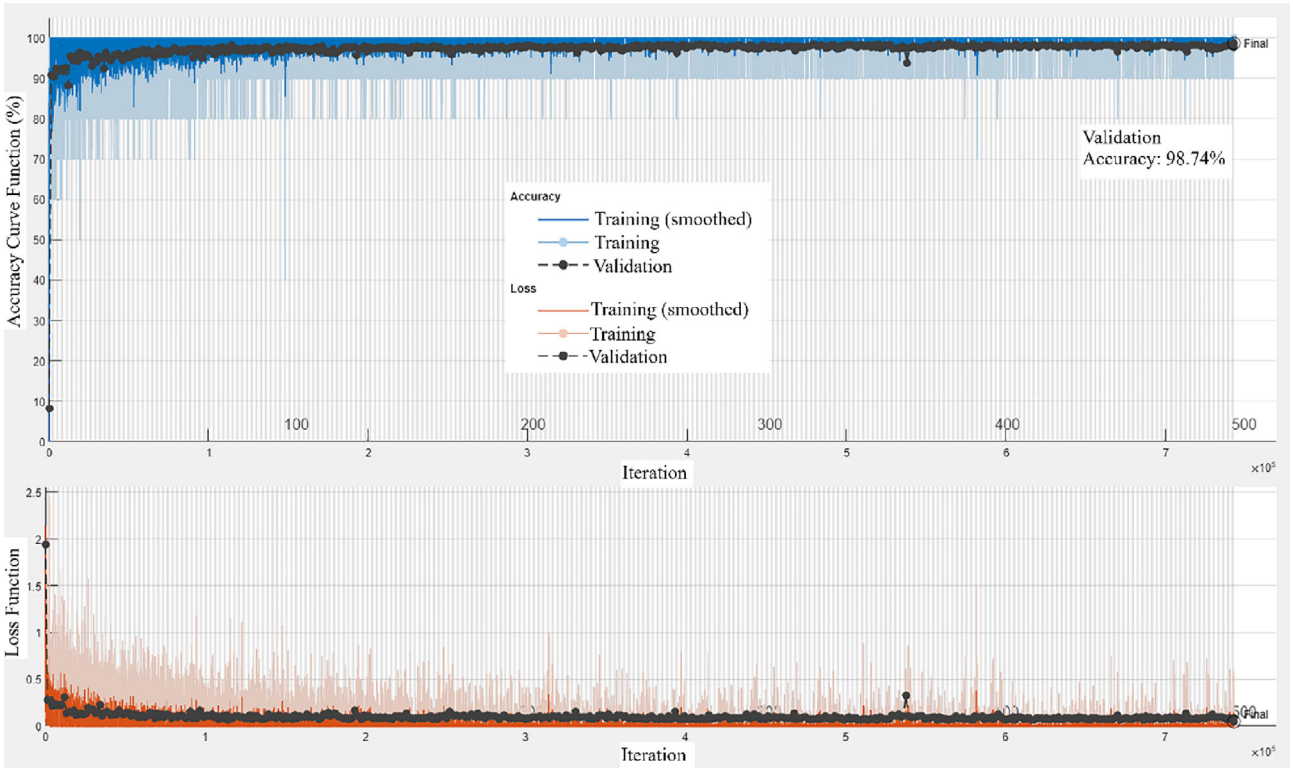

FIGURE 3 Training progress of trained model for classification. [Color figure can be viewed at wileyonlinelibrary.com]

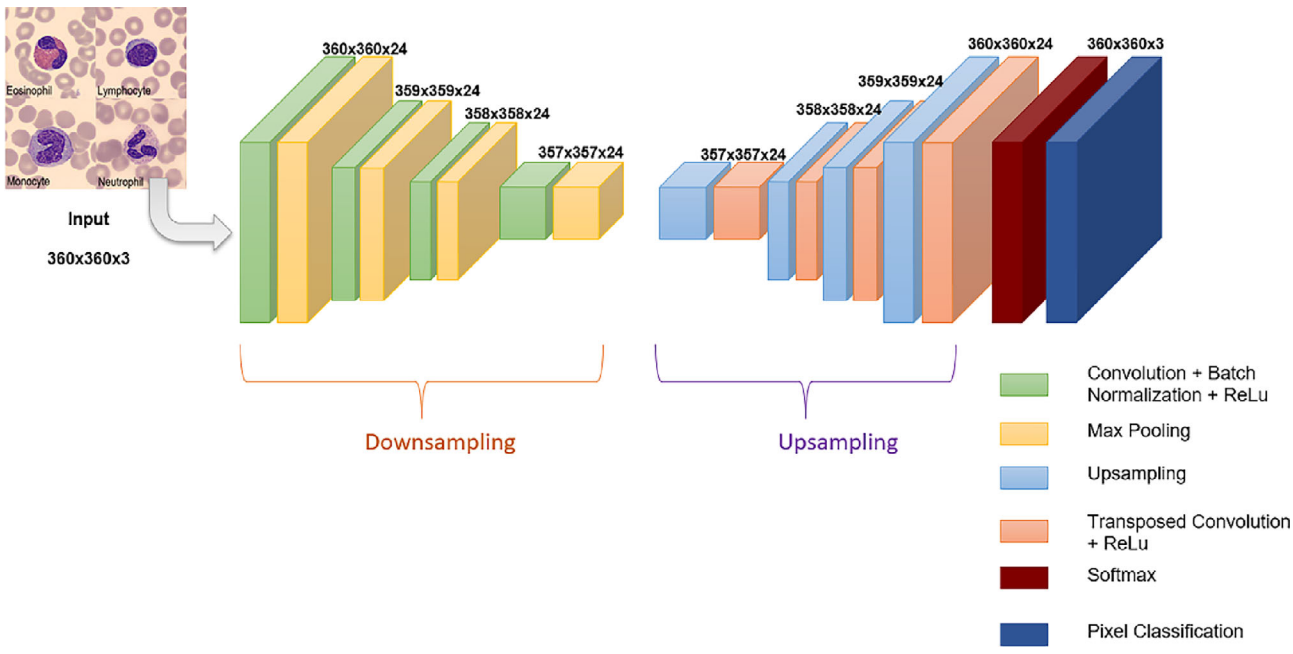

FIGURE 4 The architecture of deep learning network for semantic segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

b. UNet

The three essential components of this design are feature extraction, feature fusion, and feature reconstruction. During the process of feature extraction, the location-based feature encoder was largely employed to extract multiscale features based on convolutional blocks and residual blocks [45]. The encoder, also known as the contraction component, downsampled the input and extracted characteristics much like a CNN would. The decoder, or expanding component, upsampled and transformed the features back to a format comparable

to the original input. The UNet additionally had a concatenation part in addition to these encoder and decoder components. Information could travel through the encoder portion without having to be transformed into features. This means that the CNN model could use the information in the features of the shallower regions as well as the knowledge it learned through being compelled to construct features in the deep part [46]. Figure 5 illustrates an example UNet structure. The image input size used in this network was 360 × 360 × 3. The number of filters was 64.

c. SegNet

A sequence of organized convolutional encoder-decoder (CED) frameworks was used to develop SegNet, and each framework transferred newly learned information to the following section [47]. The SegNet structure is shown in Figure 6. The image input size used in this network was 360 × 360 × 3. The number of filters was 64.

2.3 | Performance evaluation

The results of classification and segmentation were evaluated with several different metrics. For classification, accuracy, precision, recall, and F1-score metrics were calculated for each class and overall. The mathematical formulation of these performance measures is given below [33]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (6)$$

where true positive (TP) denotes the number of WBC types accurately identified; the number of cells correctly recognized as not being the target WBC types is expressed as true negative (TN). The number of cells wrongly detected as not the target WBC types is called False-Positive (FP), while the number of cells incorrectly identified as target WBC types is called False-Negative (FN). The abovementioned metrics' results were then compared to other findings in the literature [33, 48].

For segmentation, accuracy (global accuracy, mean accuracy), intersection-over-union (meanIoU, weightedIoU), BF score (meanBFscore), and Dice similarity coefficient were calculated for each class in the nucleus and cytoplasm. The percentage of successfully recognized pixels for all classes (nucleus and cytoplasm) is referred to as global accuracy given by Equation (5):

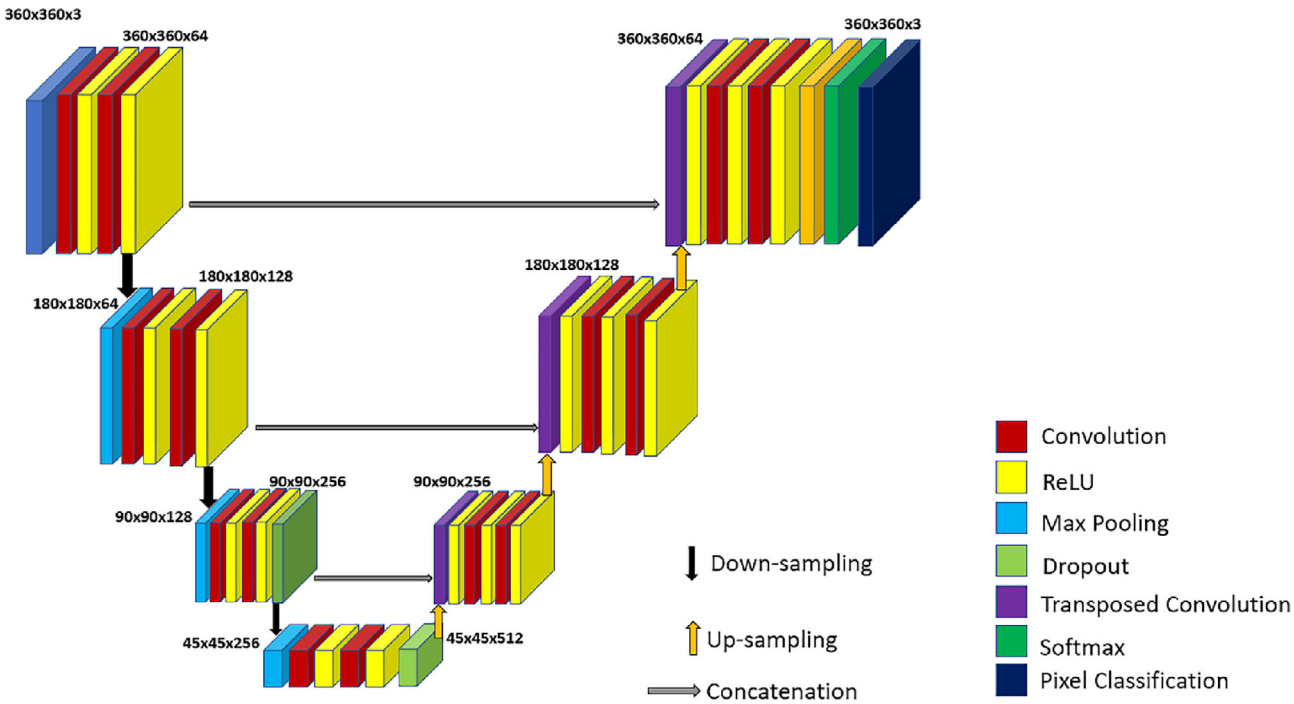

FIGURE 5 UNet structure used in the study [Color figure can be viewed at wileyonlinelibrary.com]

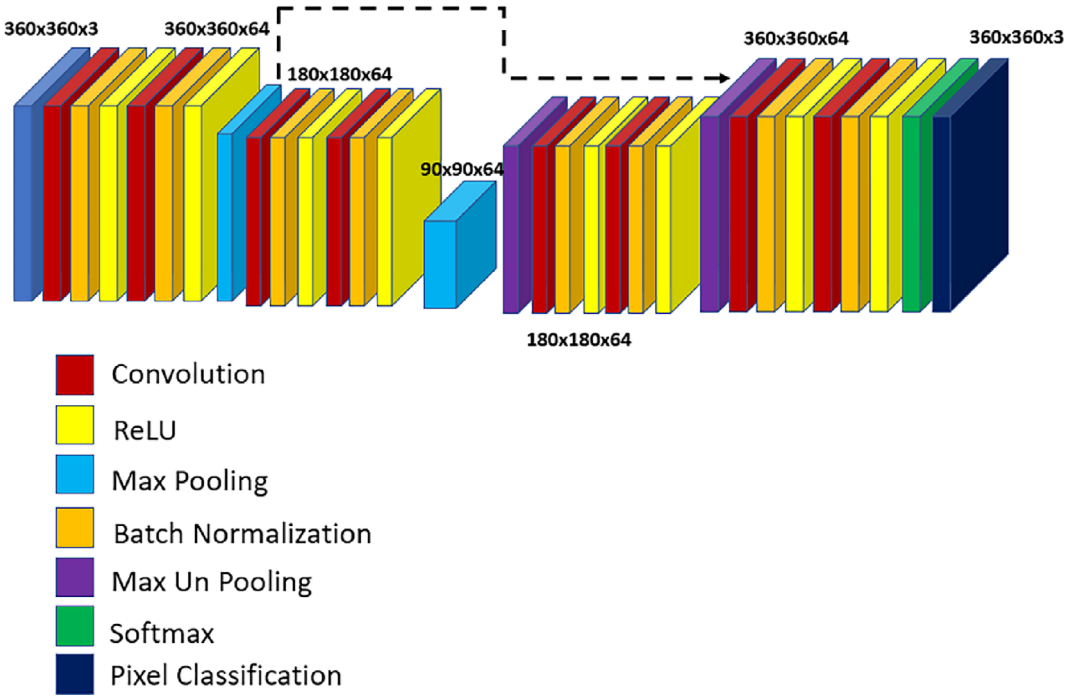

FIGURE 6 Architecture of the SegNet. [Color figure can be viewed at wileyonlinelibrary.com]

$$\text{Global Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

The global accuracy calculated for each class is called the mean accuracy, and the same formula in Equation 5 is used. The percentage of overlap score between the predicted pixel and the ground truth pixel is calculated using intersection-over-union. It is also referred to as the Jaccard index, and it determines how close the true and predicted pixels are to each other. It is determined by dividing the sum of the actual and ground truth pixel matches by the sum of the pixels in both masks (Equations 6 and 7). This value of the index ranged from 0 (no overlap) to 1 (full overlap).

$$\text{IoU} = \frac{\text{Actual} \cap \text{Ground Truth}}{\text{Actual} \cup \text{Ground Truth}} \quad (8)$$

$$\text{IoU score} = \frac{TP}{TP + FP + FN} \quad (9)$$

The quality of the segmented borders is determined by the BF score [48]. It indicates how closely each class obtained boundary fits the true boundary. Its value changes in the range [0, 1]. The BF score equals one if the contours of the predicted class and the actual data are perfectly matched. Equation 8 estimates the similarity between the predicted and ground truth images using the Dice similarity coefficient. Its value ranges from 0 to 1, with 1 denoting the closest match between the predicted and actual results:

$$\text{DSC} = 2 \times \frac{\text{Actual} \cap \text{Ground Truth}}{\text{Actual} + \text{Ground Truth}} \quad (10)$$

3 | RESULTS

In this study, CNN methods were used to classify and segment WBCs. A network of 36 layers was created for classification, and a network of 22 layers was created for segmentation. The produced CNN models were tested on the dataset and characterized according to various performance evaluation metrics.

3.1 | Performance evaluation of classification

Precision, recall, F1-score, and accuracy values of the test performance evaluation metrics for the classification model were calculated for five classes (basophils, eosinophils, monocytes, lymphocytes, and neutrophils). The CNN model created as a result of the calculations had a 98.03% balanced accuracy rate. All these calculated values are shown in Table 4.

In Figure 7, the results of the classification of WBCs are shown with the confusion matrix for the test dataset.

Precision, recall, F1-score, and accuracy values of test performance evaluation metrics were also calculated for five classes (basophils, eosinophils, monocytes, lymphocytes, and neutrophils) in the train-independent dataset, which tested the robustness of the classification model. The CNN model created as a result of the calculations had a 97.27% balanced accuracy rate. All these calculated values are shown in Table 5.

TABLE 4 Performance evaluation metrics of proposed study for classification.

Label	Precision	Sensitivity (Recall)	Specificity	F1-score	Accuracy
Basophil	0.9793	0.9726	0.9988	0.9759	0.9974
Eosinophil	0.9499	0.9843	0.9911	0.9668	0.9900
Lymphocyte	0.9733	0.9670	0.9933	0.9702	0.9871
Monocyte	0.9245	0.9437	0.9932	0.9340	0.9891
Neutrophil	0.9917	0.9814	0.9913	0.9865	0.9862
Overall	**0.9638**	**0.9698**	**0.9935**	**0.9668**	**0.9803 (Balanced)**

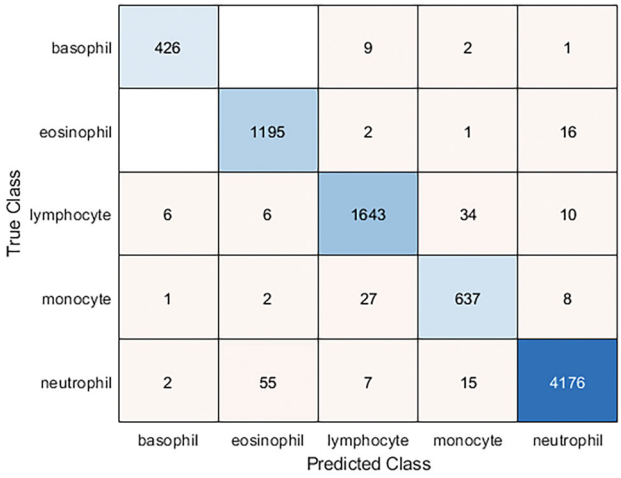

FIGURE 7 Confusion matrix representation for classification of test dataset. [Color figure can be viewed at wileyonlinelibrary.com]

In Figure 8, the results of the classification of WBCs are shown with the confusion matrix for the train-independent dataset, which was used to test the robustness of the classification model.

For both test processes, the performance evaluations of the datasets were made in their combined form. Additionally, classification accuracies were also evaluated based on the original data sets that constitute the combined data set. This evaluation was also conducted for the two combined data sets detailed above (Table 6).

As mentioned before, the performance of the classification model was examined by using the datasets individually. It was evaluated with the same performance metrics for individual use of the datasets. The CNN model created as a result of the calculations had 99.75%, 98.68% and 96.04% balanced accuracy rate for Dataset 1, Dataset 2 & 3 and Dataset 4, respectively. All these calculated values are presented in Table 7.

In order to better understand the effect of generalization by using a combined dataset, the effectiveness of trained models using a single dataset was also examined. When the datasets are used individually, the accuracy values of the trained model in classifying the data not used in the training process are presented below in Table 8.

As a final step of performance evaluation, the training and testing times were compared. The training time of the proposed CNN model was 972 min. The test time for 8367 images was 40 s. In summary, the proposed model had the capacity to process 210 images per second.

3.2 | Segmentation performance evaluation

For segmentation, global accuracy, mean accuracy, mean IoU, weighted IoU, and mean BF score values were calculated separately for each class (eosinophil, neutrophil, lymphocyte, monocyte). These values are shown in Table 9. As seen from Table 9, extensive results obtained from different deep learning methods, including custom CNN, UNet and SegNet, were examined.

Based on the performance evaluation metrics, the proposed CNN model performed better than UNet but is similar to SegNet in terms of global accuracy, mean accuracy, mean IoU, weighted IoU and mean BF score for all WBC types, as shown in Table 9. These networks are similar to each other in terms of working principles. However, there are certain differences between them, such as the number of layers, encoder depth, use of concatenation, and computational time. All these features can affect the segmentation performance for detecting different classes from the dataset, as shown in Table 9.

Moreover, to compare their performances and provide comprehensive information regarding WBC morphology, visual and numerical analyses were performed. In comparing performances, SPSS V. 22.0 (Statistical Package for Social Sciences) program for Windows was used for statistical analysis. Normal distribution of quantitative variables were analyzed with Shapiro Wilk, skewness and kurtosis values. Quantitative variables such as global accuracy, mean accuracy, mean IoU, weighted IoU and mean BF score were analyzed with Kruskal-Wallis H test when normal distribution was not suitable to compare the measurements. Significance level was accepted as $p < 0.05$. Pairwise comparisons were analyzed with Mann–Whitney U test with Bonferroni correction in case of statistical significance. Statistical significance was accepted as $p < 0.017$ for Bonferroni corrected Mann–Whitney U test.

When the statistical analysis in Table 10 is examined, statistical significance was found in the comparison of almost all performance values with UNet, but no statistical significance was found in the comparison with SegNet. Comparisons with no statistical significance are shown in Table 10.

As in classification, train-independent dataset was used for segmentation and the metric results were given in the Table 11. Subsequently, the same statistical analysis was applied to this performance data. The results obtained are similar to those in the previous performance evaluation.

When the statistical analysis in Table 12 is examined, the proposed CNN once again showed same performance compared to pre-

Label	Precision	Sensitivity (Recall)	Specificity	F1-score	Accuracy
Basophil	1.0000	0.8679	1.0000	0.9293	0.9783
Eosinophil	1.0000	0.9167	1.0000	0.9565	0.9814
Lymphocyte	0.8395	0.9577	0.9484	0.8947	0.9504
Monocyte	0.9583	0.8846	0.9926	0.9200	0.9752
Neutrophil	0.9146	1.0000	0.9718	0.9554	0.9783
Overall	**0.9425**	**0.9254**	**0,9825**	**0.9339**	**0.9727 (Balanced)**

TABLE 5 Performance evaluation metrics for evaluating robustness of classification model.

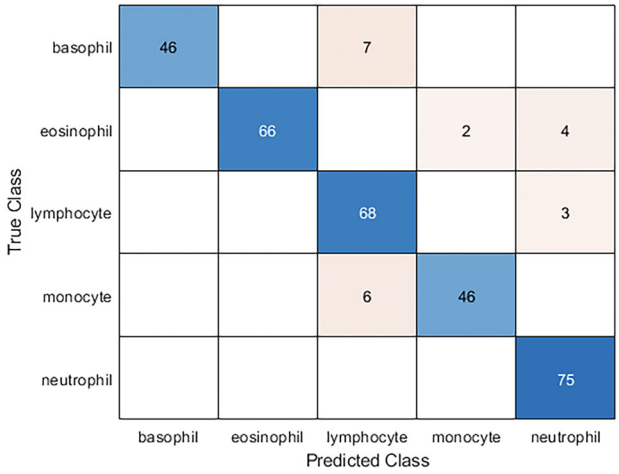

FIGURE 8 Confusion matrix representation for the train-independent dataset, which was testing the robustness of the classification model. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Examining test results separately in datasets.

	Test accuracy
Individual assessment of combined dataset	
Dataset 1	0.9937
Dataset 2 & 3	0.9735
Dataset 4	0.9247
Individual assessment of train-independent dataset	
Dataset 5	0.9902
Dataset 6	0.9552

trained networks. Statistical significance was found in the comparison of all performance values with UNet, but no statistical significance was found in the comparison with SegNet. Comparisons with no statistical significance are shown in Table 12.

In addition, accuracy, IoU, and mean BF score values were calculated for the background, nucleus, and cytoplasm parts of each cell type and are shown in Table 13. The results of different neural network structures applied for the detection of different classes, including cytoplasm, nucleus and background, in WBC cells were examined in detail.

According to Table 13, all three methods are very effective in terms of nucleus segmentation for accurate measurement in whole cells. The most successful method in this regard is UNet, but it is understood that the proposed CNN and SegNet results are very similar and are effective. The cytoplasm accuracy of the UNet method was quite low, and the other two methods were superior in performance. When we look at the comparison of the IoU ratio, it is seen that SegNet and the proposed CNN methods are at the forefront. For meanBFscore, it was observed that the most successful and efficient method on average was the recommended CNN.

When the training and testing times were compared, it was observed that the proposed CNN model performed much more effectively in both the training and testing stages. The training time for UNet was 299 min and 20 s, and the test time for 3305 images was 201 s. The training time for SegNet was 204 min and 13 s, while the test time for the same number of images was 234 s. The CNN model applied in this study was found to be faster than the other two models, with 108 min and 9 s for training time and 89 s for test time. Thus, more effective results were obtained with a simpler design in terms of evaluation metrics and processing times.

The Ground Truth Labeler tool was used for segmentation of WBCs, and images were divided into 3 sublabels: nucleus, cytoplasm, and background. The original cell images, the images tagged with the Ground Truth Labeler, the images predicted by the CNN model, and the colored representations of these labels are shown in Figure 9. Semantic segmentation metrics were employed to evaluate the test images, as indicated in Table 13. These images were used for image analysis as well, and the results were displayed as segmented images with a comparison of the anticipated labels to the ground truth labels (Figure 9).

According to Figure 9, original images, predicted labels with Dice scores, and overlaid images of predicted labels and original images are given. In terms of evaluating these results visually, it can be interpreted from Figure 9B with Dice scores that segmentation of the WBCs by using ground truth labels was quite successful. Additionally, in Figure 9C, areas detected by semantic segmentation are shown with labels and color bars of labels.

4 | DISCUSSION

This study presents the classification and segmentation of WBCs, including neutrophils, eosinophils, lymphocytes, monocytes, and basophils (not in segmentation), with convolutional neural networks. First, a deep learning approach for classification was implemented by

TABLE 7 Test results on separately trained models of the datasets.

Dataset	Precision	Sensitivity (Recall)	Specificity	F1-score	Accuracy
Dataset 1	0.9962	0.9969	0.9994	0.9966	0.9975
Dataset 2 & 3	0.9734	0.9716	0.9956	0.9725	0.9868
Dataset 4	0.9619	0.9602	0.9887	0.9605	0.9604

TABLE 8 Testing train-independent dataset with separately trained models.

Training source	Test accuracy for train-independent dataset
Dataset 1	0.3429
Dataset 2 & 3	0.4114
Dataset 4	0.4029

adjusting these blood cell images as training, validation, and test images. Second, using the same dataset, a neural network design was made for the semantic segmentation application that includes downsampling and upsampling layers.

Regarding the evaluation of the classifications, one of the most important aspects of our study was to provide a combination of four separate datasets. The training and test phases were evaluated with accuracy along with other classification parameters, such as recall, precision, and the F1 measure. The balanced accuracy of the proposed method was determined to be 98.03%, and the overall ratios of sensitivity (recall), specificity, precision, and F1 score were calculated to be 96.98%, 99.35%, 96.38%, and 96.68%, respectively. Moreover, the proposed CNN model architecture was also tested on the train-independent WBC image datasets mentioned before, and the results were given in the previous section. Since it is crucial to detect the correct cell type from any independent image source, our approach stands out from the studies in the literature. The outcome of the train-independent dataset, which consists of a mixture of two different WBC image sets, was calculated as 97.27%. In addition, other metrics, such as overall sensitivity (recall), specificity, precision, and F1 score, were also determined to be 92.54%, 98.25%, 94.25%, and 93.39% respectively. These results show that the proposed study could also work with high accuracy with train-independent datasets. Therefore, our approach can be applied to most WBC image datasets taken from any data source in the future. Moreover, the classification performance was evaluated in our study in terms of granulocytes, and an evaluation can be made through confusion matrices. When the confusion matrices were examined (Figures 7 and 8), the WBC cells with multiple granulocytes (neutrophils and eosinophils) and the cells with monogranulocytes (monocytes, lymphocytes and basophils) were similar to each other, and the proposed system confused these classes the most, as in other studies. To improve the results obtained on the basis of classes, it is thought that the sensitivity rate will be increased by using a more balanced dataset containing more new data (instead of altered data) during the training phase.

The effective success of the model evaluated in the study on the datasets was also examined separately for the datasets. When the mixed datasets were examined by separating their own images, accuracy rates of 99.37% for Dataset 1, 97.35% for Dataset 2 & 3, 92.47% for Dataset 4, 99.02% for Dataset 5, and 95.52% for Dataset 6 were obtained. It has been determined that the accuracy rate is high in other datasets except Dataset 4, which has relatively lower accuracy due to the smaller number of tested data compared to other datasets. In order to examine the performance of the model presented in the study in more detail, training and test procedures were carried out using Dataset 1, 2 & 3 and 4, separately. Satisfactory accuracy rates were obtained with values of 99.75% for Dataset 1, 98.68% for Dataset 2 & 3 and 96.04% for Dataset 4. In order to better examine the generalization effect, which is the main motivation of the study, train-independent datasets were tested using models trained with separate datasets. Accuracy rates of 34.29% for Dataset 1, 41.14% for Dataset 2 & 3 and 40.29% for Dataset 4 were obtained. The results obtained show that, with the approach presented in the study, combined training data obtained from different sources has a positive effect on generalization and increases accuracy.

In the segmentation part, WBCs in the same image dataset, excluding basophils, were used. Because of the structural differences of basophil images and the indistinct separation of cytoplasm and nucleus, these images were not included in the segmentation application. As in the classification phase, all WBCs, including neutrophils, eosinophils, lymphocytes, and monocytes, were segmented through the semantic segmentation network, including the proposed CNN with well-known pretrained UNet and SegNet networks, and as mentioned in the results section, evaluation criteria were calculated separately for each cell. In addition to semantic segmentation metrics, segmented images of WBCs were obtained by calculating the Dice index values for the proposed CNN. The Dice scores of each cell were calculated separately for the cytoplasm and the nucleus, and their visuals are shown in the results section. In addition, images were added by calculating the Dice index ratios in the images where the cytoplasm and nucleus classes were overlaid together. When the segmentation results are examined, the results of the nuclei segmentation of four different WBCs are calculated for eosinophils, neutrophils, lymphocytes, and monocytes as 97.78%, 98.45%, 96.89% and 96.84%, respectively. In addition, statistical analyses of three different network performances used for segmentation were performed to detect any significant differences. Accordingly, when SegNet and CNN are considered, no significant difference was observed in most of the performance metrics. On the other hand, it was found that there is a significant difference between proposed CNN and UNet in terms of performance metrics. The test results carried out to examine the robustness effect, which is the main motivation of the study, have

TABLE 9 Performance evaluation metrics of proposed study and pre-trained networks for semantic segmentation.

	Label	Global accuracy	Mean accuracy	Mean IoU	Weighted IoU	Mean BF score
Proposed CNN	Eosinophil	0.9904	0.9713	0.9369	0.9815	0.9447
	Lymphocyte	0.9920	0.9181	0.8443	0.9857	0.9073
	Monocyte	0.9825	0.9526	0.9083	0.9669	0.8818
	Neutrophil	0.9900	0.9669	0.9406	0.9805	0.9402
UNet	Eosinophil	0.9748	0.8947	0.7902	0.9561	0.7368
	Lymphocyte	0.9904	0.8197	0.7696	0.9823	0.8722
	Monocyte	0.9729	0.8953	0.8359	0.9490	0.7946
	Neutrophil	0.9800	0.9192	0.8453	0.9628	0.8192
SegNet	Eosinophil	0.9897	0.9679	0.9310	0.9802	0.9423
	Lymphocyte	0.9916	0.9006	0.8353	0.9848	0.9066
	Monocyte	0.9813	0.9430	0.9014	0.9645	0.8677
	Neutrophil	0.9894	0.9644	0.9370	0.9794	0.9359

TABLE 10 Statistical analysis of the performances with their pairwise comparisons of segmentation methods.

		p-values			
	Method	Eosinophil	Lymphocyte	Monocyte	Neutrophil
Global Accuracy	Proposed CNN vs. SegNet	$p = 0.072$	$p = 0.027$	$p = 0.348$	$p = 0.187$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$p = 0.127$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Mean Accuracy	Proposed CNN vs. SegNet	$p = 0.052$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$p = 0.246$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Mean IoU	Proposed CNN vs. SegNet	$p = 0.084$	$p = 0.036$	$p = 0.045$	$p = 0.056$
	Proposed CNN vs. UNet	$\mathbf{p = 0.007}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Weighted IoU	Proposed CNN vs. SegNet	$p = 0.054$	$p = 0.253$	$p = 0.245$	$p = 0.142$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Mean BFScore	Proposed CNN vs. SegNet	$p = 0.103$	$p = 0.159$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$

*Refers significance for $p < 0.017$; All comparisons for methods show statistically significant difference as $p < 0.001$.

similar results. Although significant differences were detected with UNet and proposed CNN in the evaluation of segmentation results using performance metrics, the same situation did not occur for Seg-Net and proposed CNN. According to the results obtained, a model structure that can produce better results than the UNet model, which is frequently used in the literature, and can also compete with the Seg-Net model, has been obtained. On the other hand, the proposed CNN also has an advantage in terms of computational time, which is shorter than that of other pretrained neural networks. Therefore, it can be said that it provides a time-efficient and resourceful option compared to UNet and SegNet.

As shown in Tables 14 and 15, the results of our study and some selected studies were reported in regard to classification and segmentation. The main criterion for the selected studies is that they are literature studies related to the datasets included in our study. However, the proposed methods have a unique characteristic that differs from these other studies because none of the related studies combined and used more than two datasets in the classification stage. In this study, as mentioned in the previous sections, a mixture of four different datasets was used as the training set to classify and segment WBCs. Furthermore, to evaluate the robustness of the classification model and segmentation model, an independent dataset that was unrelated to the training phase was evaluated, and the results are given in Table 5 and Table 11. In regard to classification and segmentation, as stated, considering the innovative side of our approach, a logical comparison could not be made since there is no similar approach among literature studies. The overall accuracy rate in the literature using a single dataset varied between 83.44% [19] and 99.73% [33] for classification and 97.40% [30] and 99.42% [32] for segmentation in reported studies. Both our multidataset rate (overall accuracy) and

TABLE 11 Performance evaluation metrics of proposed study and pre-trained networks for semantic segmentation with train-independent dataset.

	Label	Global accuracy	Mean accuracy	Mean IoU	Weighted IoU	Mean BF score
Proposed CNN	Eosinophil	0.9294	0.9296	0.9156	0.9597	0.9233
	Lymphocyte	0.9350	0.8891	0.8904	0.9156	0.8784
	Monocyte	0.9292	0.9307	0.8869	0.9448	0.8688
	Neutrophil	0.9194	0.9472	0.9114	0.9505	0.9110
UNet	Eosinophil	0.8904	0.8744	0.8710	0.9152	0.8781
	Lymphocyte	0.9082	0.7975	0.8480	0.8981	0.8493
	Monocyte	0.9029	0.8605	0.8028	0.9128	0.8126
	Neutrophil	0.8858	0.8881	0.8558	0.9307	0.8803
SegNet	Eosinophil	0.9275	0.9231	0.9067	0.9553	0.9209
	Lymphocyte	0.9366	0.8733	0.8989	0.9163	0.8792
	Monocyte	0.9201	0.9226	0.8719	0.9337	0.8599
	Neutrophil	0.9105	0.9438	0.9167	0.9586	0.9156

TABLE 12 Statistical analysis of the robustness performances with their pairwise comparisons of segmentation methods.

		p-values			
	Method	Eosinophil	Lymphocyte	Monocyte	Neutrophil
Global Accuracy	Proposed CNN vs. SegNet	$p = 0.027$	$p = 0.127$	$p = 0.348$	$p = 0.187$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p = 0.005}^*$
Mean Accuracy	Proposed CNN vs. SegNet	$p = 0.105$	$\mathbf{p < 0.001}^*$	$p = 0.078$	$p = 0.246$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Mean IoU	Proposed CNN vs. SegNet	$p = 0.058$	$p = 0.345$	$\mathbf{p = 0.004}^*$	$p = 0.056$
	Proposed CNN vs. UNet	$\mathbf{p = 0.007}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$
Weighted IoU	Proposed CNN vs. SegNet	$p = 0.072$	$p = 0.253$	$\mathbf{p = 0.007}^*$	$p = 0.085$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$	$p = 0.154$	$\mathbf{p < 0.001}^*$
Mean BFScore	Proposed CNN vs. SegNet	$p = 0.046$	$p = 0.036$	$p = 0.055$	$p = 0.054$
	Proposed CNN vs. UNet	$\mathbf{p < 0.001}^*$	$\mathbf{p = 0.003}^*$	$\mathbf{p < 0.001}^*$	$\mathbf{p < 0.001}^*$

*Refers significance for $p < 0.017$; All comparisons for methods show statistically significant difference as $p < 0.001$.

robustness evaluation rate (overall accuracy) were within the acceptable successful limits considering these studies to make a relative comparison. It was also observed that the presented study was slightly more successful than the only study using multiple datasets [34]. However, since no study was found that tested a dataset that was not used in the training phase, an accurate comparison could not be made as we mentioned before. In the study, the effectiveness of the use of data obtained from different sources was determined as a previously stated situation in results section. Datasets were also evaluated for the effectiveness of the model used in the study. When Table 14 is examined, it will be seen that the classification model has test results with accuracy that can compete with other classification studies, especially in the CellaVision dataset studies.

Furthermore, our study stands out from most of the studies because it also segments WBCs with different deep learning approaches. Most of the related studies consist of either classification or segmentation. Two of the studies on segmentation were executed by using four different datasets separately [31, 32]. However, in these studies and others, the combined use of datasets was not found in the literature. Again, although not an exact comparison, the accuracy rates of the presented work and the segmentation studies whose results are summarized in Table 15 have a satisfactory similarity. Considering our segmentation process, it has been seen that the presented approach is effective for segmentation as well as for classification. As seen in the Table 15, in the test results performed with the independent dataset, our approach resulted in high accuracy in segmentation in addition to classification. In this sense, an accuracy of 92.82% was achieved and it can be said that the system works very effectively with the independent dataset. As a result, a detailed analysis including the classification and

TABLE 13 Performance evaluation metrics for the nucleus, cytoplasm, and background parts of each class.

	Label	Eosinophil			Lymphocyte			Monocyte			Neutrophil		
	Metrics	Background	Nucleus	Cytoplasm	Background	Nucleus	Cytoplasm	Background	Nucleus	Cytoplasm	Background	Nucleus	Cytoplasm
Proposed CNN	Accuracy	0.9952	0.9778	0.9411	0.9963	0.9689	0.7890	0.9916	0.9684	0.8978	0.9964	0.9845	0.9200
	IoU	0.9922	0.9340	0.8846	0.9946	0.9288	0.6095	0.9849	0.9331	0.8069	0.9908	0.9516	0.8793
	Mean BF score	0.9702	0.9382	0.9256	0.9758	0.8839	0.8617	0.9338	0.8621	0.8495	0.9601	0.9558	0.9047
UNet	Accuracy	0.9974	09976	0.6890	0.9985	0.9908	0.4698	0.9963	0.9931	0.6965	0.9982	0.9970	0.7625
	IoU	0.9936	0.7032	0.5054	0.9963	0.8814	0.4312	0.9837	0.8549	0.6689	0.9899	0.7971	0.7491
	Mean BF score	0.9776	0.5054	0.7274	0.9867	0.7748	0.8530	0.9211	0.6976	0.7650	0.9522	0.7259	0.7793
SegNet	Accuracy	0.9952	0.9734	0.9353	0.9967	0.9723	0.9770	0.9929	0.9607	0.8755	0.9962	0.9803	0.9166
	IoU	09919	0.9222	0.8789	0.9944	0.9220	0.8857	0.9837	0.9242	0.7963	0.9903	0.9465	0.8743
	Mean BF score	0.9711	0.9288	0.9270	0.9770	0.5897	0.8559	0.9247	0.8425	0.8360	0.9552	0.9531	0.8993

segmentation of WBCs was performed in this study. With this feature, our study stands out because it includes both classification and segmentation processes.

In the present study, it was observed that satisfactory results were produced in terms of success rates, considering that the complexity increases with the mixed use of data obtained from multiple sources and that the analysis becomes difficult. In most of the studies mentioned in the literature summary, models were built with approaches (transfer learning, fine-tuning, etc.) that include pretrained networks [4–27]. All these models are more complex than our proposed networks. Therefore, considering the complexity and learnable parameters of pretrained networks, the time required to complete both training and testing processes is expected to be longer than the presented work. In addition, the studies that created their own network did not specify the number of images processed per second, so no comparison could be made regarding time durations. Moreover, the complexity of the proposed networks increases the potential to process large-scale data with hardware that is considered average today. For this reason, the proposed networks would be a very suitable candidate for use in CAD system-based automated equipment, as they are efficient in terms of quick data processing and low computational cost.

5 | CONCLUSION

In this study, a deep learning approach for the classification and segmentation of WBCs was presented. The current study is distinguished from other studies due to the utilization of a mixture of four different datasets with considerable classification accuracy. Additionally, to test the robustness of the network proposed in the study, studies were also conducted with a completely independent dataset that was not used in training. In addition, a more detailed analysis was presented with the segmentation technique applied to WBCs with the same mixture by utilizing custom and pretrained networks. With this study, both the classification and segmentation of WBCs were performed with high accuracy (98.03% and 98.87% overall accuracy), contributing to the studies in this field with a new perspective of multisource datasets. Moreover, promising results for both classification and segmentation (97.27% and 92.82% overall accuracy, respectively) were obtained from robustness evaluation tests of the proposed network using different datasets than the training data. Finally, in the experiments conducted on the proposed CNN architecture and some well-known networks in our segmentation study, it was found that our CNN model can compete with other networks (especially SegNet) and is successful (98,87% overall accuracy) in terms of most measured parameters. Furthermore, segmentation test results performed with the independent dataset also reached an accuracy rate of 92.82%. In addition to the successful performance results, the ability of the network proposed in this study to produce fast results and use resources efficiently, especially for applications where large amounts of data need to be processed, highlights its usability in automated systems.

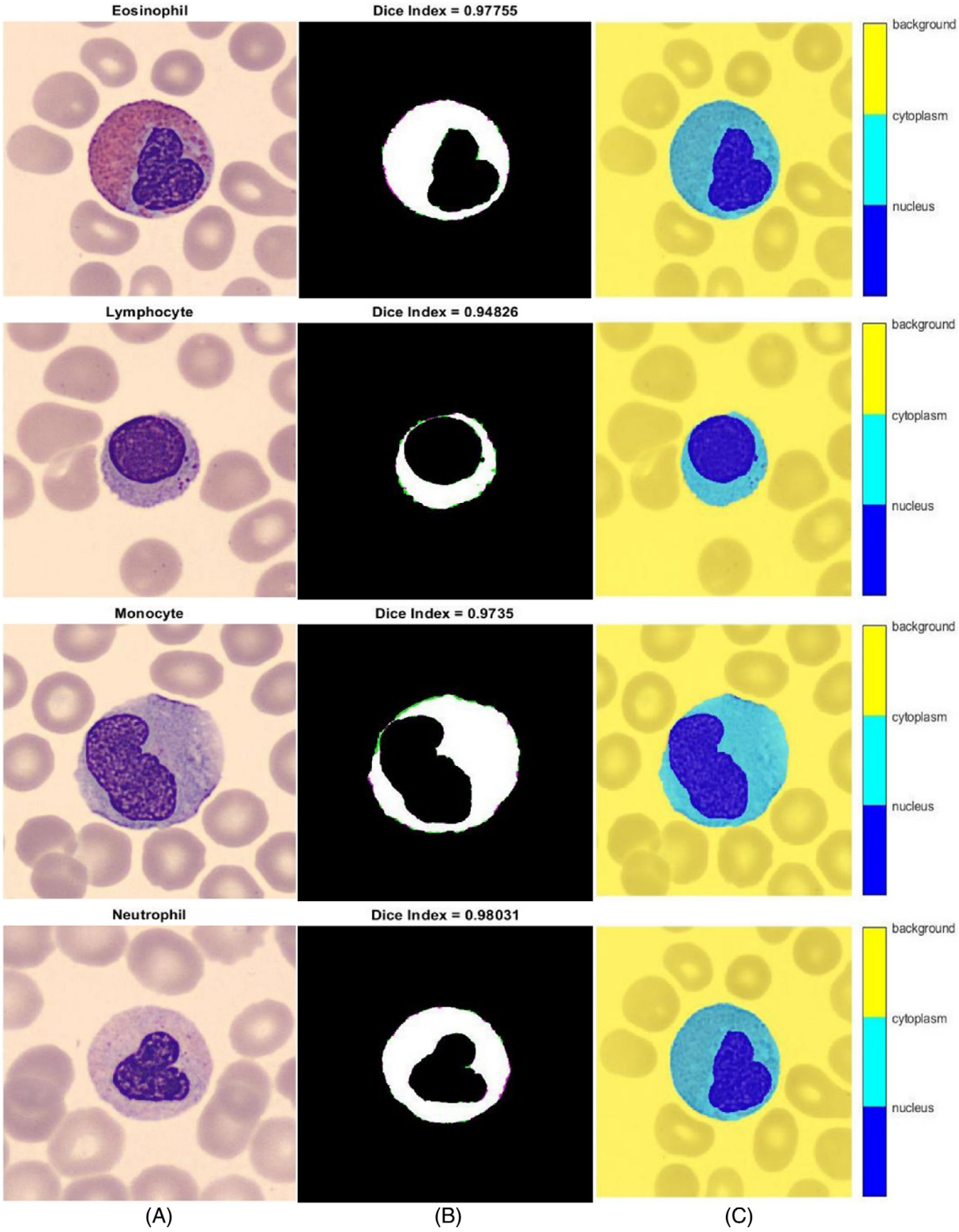

FIGURE 9 Results of white blood cells (WBC) segmentation. (A) original images, (B) predicted labels on ground truth labels with dice scores, (C) segmentation results with color bar. [Color figure can be viewed at wileyonlinelibrary.com]

AUTHOR CONTRIBUTIONS

Gökay Karayeğen: Conceptualization; investigation; methodology; validation; visualization; software; formal analysis; supervision. **Şeyma Nur Özcan:** Investigation; software; formal analysis. **Tansel Uyar:** Conceptualization; investigation; methodology; visualization; software; formal analysis; validation; supervision.

TABLE 14 Comparison between proposed study and related studies for classification.

		Cellavision	[10]	[11]	[13]	[17]	
			Merino et al.: 96%	Shaheen et al.: 98.58%	Bagido et al.: 96.21%	Gavas et al.: 99.51%	
Studies for classification (Overall accuracy)	Literature	Kaggle	[4]	[9]	[15]	[19]	[20]
			Ridoy et al.: 95.97%	Siddique et al.: 93.80%	Toğaçar et al.: 97.95%	Yildirim et al.: 83.44%	Mohamed et al.: 92%
			[21]	[22]	[23]	[24]	[26]
			Patil et al.: 95.89%	Khan et al.: 99.12%	Praveen et al.: 90%	Toğaçar et al.: 97.78%	Girdhar et al.: 98.55%
		Multi dataset	Çınar et al. [33] Kaggle: 99.73% LISC: Not Relevant	Long et al. [28] CellaVision: 97.29% Kaggle: 99.47% ALL-IDB2: Not Relevant	Yao et al. [29] Kaggle: 91.6% Private Dataset: Not Relevant	Kutlu et al. [34] Kaggle and LISC: 97.52%	
	Proposed Study		CellaVision: 99.75% Raabin-WBC: 98.68% Kaggle: 96.04% CellaVision, Raabin-WBC and Kaggle: 98.03%				
Robustness evaluation for classification (Test with train-independent dataset) (Overall accuracy)	Literature		N/A				
	Proposed Study		LISC, BCD: 97.27%				

TABLE 15 Comparison between proposed study and related studies for segmentation.

Studies for segmentation (overall accuracy)	Literature	Datasets	Mohammadi et al. [27]	Roy et al. [30]	Lu et al. [31]	Banik et al. [32]	
				CellaVision: 97.40%	CellaVision: 98.33%	CellaVision: 98.86%	Kadry et al. [47] LISC: 97.73%
			CellaVision: 95% (nucleus), 91% (cytoplasm)	LISC: 98.22%	LISC: 94.20%	Kaggle: 99.42%	
					JTSC: Not relevant	ALL-IDB2: Not relevant	
				JTSC: Not relevant	Private dataset: Not relevant	JTSC: Not relevant	
	Proposed Study		CellaVision, Raabin-WBC and Kaggle: 98.87%				
Robustness Evaluation for Segmentation (Test with Train-Independent Dataset) (Overall Accuracy)	Literature		N/A				
	Proposed Study		LISC, BCD: 92.82%				

ACKNOWLEDGMENTS

We would like to thank Dr. Didem Sakaryalı Uyar for her contribution to the statistical evaluation of the segmentation results.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cyto.a.24839.

REFERENCES

1. Turgeon ML. In: Turgeon ML, editor. Clinical hematology: Theory & procedures. 5th ed. Philadelphia: Lippincott Williams & Wilkins; 2012.
2. Pozdnyakova O, Connell NT, Battinelli EM, Connors JM, Fell G, Kim AS. Clinical significance of CBC and WBC morphology in the diagnosis and clinical course of COVID-19 infection. Am J Clin Pathol. 2021;155:364–75. Available at: https://academic.oup.com/ajcp/article/155/3/364/6017543
3. Thachil J, Bates I. Approach to the diagnosis and classification of blood cell disorders. Dacie and Lewis Practical Haematology. Elsevier; 2017. p. 497–510. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780702066962000230
4. Ridoy AR, Islam R. A lightweight convolutional neural network for white blood cells classification. 2020 23rd international conference on computer and information technology (ICCIT). IEEE; 2020. p. 1–5. Available at: https://ieeexplore.ieee.org/document/9392649/
5. Aziz S, Bilal M, Khan MU, Amjad F. Deep learning-based automatic morphological classification of leukocytes using blood smears. 2020 international conference on electrical, communication, and computer engineering (ICECCE). IEEE; 2020. p. 1–5. Available at: https://ieeexplore.ieee.org/document/9179246/
6. Hartanto CA, Kurniawan S, Arianto D, Arymurthy AM. DCGAN-generated synthetic images effect on white blood cell classification. IOP Conf Ser Mater Sci Eng. 2021;1077:012033 Available at: https://iopscience.iop.org/article/10.1088/1757-899X/1077/1/012033
7. Almezhghwi K, Serte S. Improved classification of white blood cells with the generative adversarial network and deep convolutional neural network. Comput Intell Neurosci. 2020;2020:1–12. Available at: https://www.hindawi.com/journals/cin/2020/6490479/
8. Ma L, Shuai R, Ran X, Liu W, Ye C. Combining DC-GAN with ResNet for blood cell image classification. Med Biol Eng Comput. 2020;58:1251–64. Available at: http://link.springer.com/10.1007/s11517-020-02163-3
9. Siddique MAI, Bin AAZ, Matin A. An improved deep learning based classification of human white blood cell images. 2020 11th international conference on electrical and computer engineering (ICECE). IEEE; 2020. p. 149–152. Available at: https://ieeexplore.ieee.org/document/9393156/
10. Merino A, Vlagea A, Molina A, Egri N, Laguna J, Barrera K, et al. Atypical lymphoid cells circulating in blood in COVID-19 infection: morphology, immunophenotype and prognosis value. J Clin Pathol. 2022;75:104–11. Available at: https://jcp.bmj.com/lookup/doi/10.1136/jclinpath-2020-207087
11. Shaheen M, Khan R, Biswal RR, Ullah M, Khan A, Uddin MI, et al. Acute myeloid leukemia (AML) detection using AlexNet model. Complexity. 2021;2021:1–8. Available at: https://www.hindawi.com/journals/complexity/2021/6658192/
12. Macawile MJ, Quinones VV, Ballado A, Dela CJ, Caya MV. White blood cell classification and counting using convolutional neural network. 2018 3rd international conference on control and robotics engineering (ICCRE). IEEE; 2018. p. 259–263. Available at: https://ieeexplore.ieee.org/document/8376476/
13. Bagido RA, Alzahrani M, Arif M. White blood cell types classification using deep learning models. IJCSNS Int J Comp Sci Netw Secur. 2021;21:223 Available at: https://doi.org/10.22937/IJCSNS.2021.21.9.30
14. Khouani A, El Habib DM, Mahmoudi SA, Chikh MA, Benzineb B. Automated recognition of white blood cells using deep learning. Biomed Eng Lett. 2020;10:359–67. Available at: https://link.springer.com/10.1007/s13534-020-00168-3
15. Toğaçar M, Ergen B, Cömert Z. Classification of white blood cells using deep features obtained from Convolutional Neural Network models based on the combination of feature selection methods. Appl Soft Comput. 2020;97:106810 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1568494620307481
16. Reena MR, Ameer PM. Localization and recognition of leukocytes in peripheral blood: a deep learning approach. Comput Biol Med. 2020;126:104034 Available at: https://linkinghub.elsevier.com/retrieve/pii/S0010482520303656
17. Gavas E, Olpadkar K. Deep CNNs for peripheral blood cell classification. 2021 Available at: http://arxiv.org/abs/2110.09508
18. Hegde RB, Prasad K, Hebbar H, Singh BMK. Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. Biocybern Biomed Eng. 2019;39:382–92. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0208521618304819
19. Yildirim M, Çinar A. Classification of white blood cells by deep learning methods for diagnosing disease. Revue d'Intelligence Artificielle. 2019;33:335–40. Available at: http://www.iieta.org/journals/ria/paper/10.18280/ria.330502
20. Mohamed EH, El-Behaidy WH, Khoriba G, Li J. Improved white blood cells classification based on pre-trained deep learning models. J Commun Softw Syst. 2020;16:37–45. Available at: https://jcoms.fesb.unist.hr/10.24138/jcomss.v16i1.818/
21. Patil AM, Patil MD, Birajdar GK. White blood cells image classification using deep learning with canonical correlation analysis. IRBM. 2021;42:378–89. Available at: https://linkinghub.elsevier.com/retrieve/pii/S195903182030141X
22. Khan A, Eker A, Chefranov A, Demirel H. White blood cell type identification using multi-layer convolutional features with an extreme-learning machine. Biomed Signal Proc Contr. 2021;69:102932 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1746809421005292
23. Praveen N, Punn NS, Sonbhadra SK, Agarwal S, Syafrullah M, Adiyarta K. White blood cell subtype detection and classification. 2021 8th international conference on electrical engineering, computer science and informatics (EECSI). Vol 2021-October. IEEE; 2021. p. 203–207. Available at: https://ieeexplore.ieee.org/document/9624268/
24. Togacar M, Ergen B, Sertkaya ME. Subclass separation of white blood cell images using convolutional neural network models. Elektronika ir Elektrotechnika. 2019;25:63–8. Available at: http://eejournal.ktu.lt/index.php/elt/article/view/24358
25. Shahin AI, Guo Y, Amin KM, Sharawi AA. White blood cells identification system based on convolutional deep neural learning networks. Comput Methods Prog Biomed. 2019;168:69–80. Available at: https://linkinghub.elsevier.com/retrieve/pii/S016926071730411X
26. Girdhar A, Kapur H, Kumar V. Classification of white blood cell using convolution neural network. Biomed Signal Proc Contr. 2022;71:103156 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1746809421007539
27. Mohammadi E, Orooji M. An unsupervised and supervised combined approach for white blood cells segmentation. 2018 25th national and 3rd international Iranian conference on biomedical engineering (ICBME). IEEE; 2018. p. 1–6. Available at: https://ieeexplore.ieee.org/document/8703561/

28. Long F, Peng J-J, Song W, Xia X, Sang J. BloodCaps: a capsule network based model for the multiclassification of human peripheral blood cells. Comput Methods Prog Biomed. 2021;202:105972 Available at: https://linkinghub.elsevier.com/retrieve/pii/S016926072100047X
29. Yao X, Sun K, Bu X, Zhao C, Jin Y. Classification of white blood cells using weighted optimized deformable convolutional neural networks. Artif Cells Nanomed Biotechnol. 2021;49:147–55. Available at: https://www.tandfonline.com/doi/full/10.1080/21691401.2021.1879823
30. Roy RM, Ameer PM. Segmentation of leukocyte by semantic segmentation model: a deep learning approach. Biomed Signal Proc Contr. 2021;65:102385 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1746809420304912
31. Lu Y, Qin X, Fan H, Lai T, Li Z. WBC-net: a white blood cell segmentation network based on UNet++ and ResNet. Appl Soft Comput. 2021;101:107006 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1568494620309455
32. Banik PP, Saha R, Kim K-D. An automatic nucleus segmentation and CNN model based classification method of white blood cell. Expert Syst Appl. 2020;149:113211 Available at: https://linkinghub.elsevier.com/retrieve/pii/S0957417420300373
33. Çınar A, Tuncer SA. Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM. SN Appl Sci. 2021;3:503 Available at: https://link.springer.com/10.1007/s42452-021-04485-9
34. Kutlu H, Avci E, Özyurt F. White blood cells detection and classification based on regional convolutional neural networks. Med Hypotheses. 2020;135:109472 Available at: https://linkinghub.elsevier.com/retrieve/pii/S0306987719310680
35. AL-Dulaimi K, Banks J, Nugyen K, Al-Sabaawi A, Tomeo-Reyes I, Chandran V. Segmentation of white blood cell, nucleus and cytoplasm in digital haematology microscope images: a review–challenges, current and future potential techniques. IEEE Rev Biomed Eng. 2021;14:290–306. Available at: https://ieeexplore.ieee.org/document/9124647/
36. Basu A, Senapati P, Deb M, Rai R, Dhal KG. A survey on recent trends in deep learning for nucleus segmentation from histopathology images. Evol Syst. 2023. Available at: https://link.springer.com/10.1007/s12530-023-09491-3;15:203–48.
37. Acevedo A, Merino A, Alférez S, Molina Á, Boldú L, Rodellar J. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data Brief. 2020;30:105474 Available at: https://linkinghub.elsevier.com/retrieve/pii/S2352340920303681
38. Kouzehkanan ZM, Saghari S, Tavakoli S, Rostami P, Abaszadeh M, Mirzadeh F, et al. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. Sci Rep. 2022;12:1123 Available at: https://www.nature.com/articles/s41598-021-04426-x
39. Chen N. BCCD (blood cell count and detection) dataset is a small-scale dataset for blood cells detection. GitHub. 2018 Available at: https://github.com/Shenggan/BCCD_Dataset. Accessed February 10, 2023
40. Rezatofighi SH, Soltanian-Zadeh H. Automatic recognition of five types of white blood cells in peripheral blood. Comput Med Imaging Graph. 2011;35:333–43. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0895611111000048
41. Aslan A. WBC & RBC detection dataset from peripheral blood smears. GitHub. 2020 Available at: https://github.com/draaslan/blood-cell-detection-dataset. Accessed February 10, 2023
42. Keskar NS, Socher R. Improving generalization performance by switching from Adam to SGD. Comp Sci ArXiv. 2017 Available at: http://arxiv.org/abs/1712.07628
43. Zhou P, Feng J, Ma C, Xiong C, Hoi S, Towards EW. Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. p. 21285–21296. Available at: http://arxiv.org/abs/2010.05627 2020.
44. Gupta A, Ramanath R, Shi J, Keerthi SS. Adam vs. SGD: closing the generalization gap on image classification. OPT2021: 13th annual workshop on optimization for machine learning. 2021.
45. Alharbi AH, Aravinda CV, Lin M, Venugopala PS, Reddicherla P, Shah MA. Segmentation and classification of white blood cells using the UNet. Contrast Media Mol Imaging. 2022;2022:1–8. Available at: https://www.hindawi.com/journals/cmmi/2022/5913905/
46. Huynh HT, Dat VVT, Anh HB. White blood cell segmentation and classification using deep learning coupled with image processing technique. Communications in computer and information science: future data and security engineering. Big data, security and privacy, smart city and industry 4.0 applications. FDSE 2021. Volume 1500; 2021. p. 399–410. Available at: https://link.springer.com/10.1007/978-981-16-8062-5_27
47. Kadry S, Rajinikanth V, Taniar D, Damaševičius R, Valencia XPB. Automated segmentation of leukocyte from hematological images—a study using various CNN schemes. J Supercomput. 2022;78:6974–94. Available at: https://link.springer.com/10.1007/s11227-021-04125-4
48. Fernandez-Moral E, Martins R, Wolf D, Rives P. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. 2018 IEEE intelligent vehicles symposium (IV). IEEE; 2018. p. 1051–1056. Available at: https://ieeexplore.ieee.org/document/8500497/

How to cite this article: Özcan ŞN, Uyar T, Karayeğen G. Comprehensive data analysis of white blood cells with classification and segmentation by using deep learning approaches. Cytometry. 2024;105(7):501–20. https://doi.org/10.1002/cyto.a.24839