American Journal of Biomedical Sciences
ISSN: 1937-9080
nwpii.com/ajbms

Leukocyte (White Blood Cell) Classification with a Multi-stage Support Vector Machine

Hoang Truong Kien, Nguyen Hoai Phuong, Hoang Thi Luyen, Nguyen Minh Duc, Duong Trong Luong*

Department of Electronic Technology and Biomedical Engineering, Hanoi University of Science and Technology, Vietnam

***Corresponding Author**
Duong Trong Luong
Department of Electronic Technology and Biomedical Engineering
Hanoi University of Science and Technology
Vietnam
Email: luong.duongtrong@hust.edu.vn

Received:18 July 2020; | Revised:08 August 2020; | Accepted:01 December 2020

Abstract

In this paper we present an automatic method for leukocyte classification using support vector machines in multiple stages that results in higher accuracy and less overfitting than other automatic methods. White blood cells (WBC), also called leukocytes, play an important role in the immune system to help protect the body from virus and bacterial diseases. Leukocyte tests can help diagnose a number of diseases related to blood disorders such as acute leukemia, chronic myeloid leukemia. Manual blood cell classification is commonly used in clinics and hospitals, automated cell classification systems may assist clinicians to increase efficiency and accuracy of diagnosis. In recent years, along with the development of artificial intelligence, there have been many studies using automatic methods to classify and count leukocytes. In this paper, we propose a Multi-stage support vector machine method (Multistage SVM) to classify leukocytes into 4 classes with 93% accuracy and less overfitting than other automatic methods.

Keywords: Leukocyte, classification, Support vector machine, Multi-Stage, Blood disorder

1. Introduction

1.1 Background

Automatic blood cell classification plays an important role in diagnosing a number of blood cancers and other diseases related to blood disorders. Manually classifying and counting leukocytes from microscopes images is fraught with difficulties: (a) time consuming, (b) experience dependence, (c) fatigue when analyzing large numbers of samples, (d) dependence sample variation including leukocyte shapes, number of cells, staining method and repetition of cells during examination. Automatic cell classification could help alleviate many of these issues. In automatic cell classification the process typically includes: 1-cell segmentation, 2-feature extraction, 3-cell classification.

Cell segmentation and feature extraction play an important role in the classification of leukocytes. Accurate segmentation improves the feature indices

including cell shapes, cell nuclei, affect of heterogeneous luminosity, staining methods, and changes in cell topology in the maturation stages, image rotation, cell magnification.

In previous studies on blood cell segmentation accuracy is the performance measure. K-means clustering method to classify blood cell in color space [1], this segmentation method achieved an average accuracy of 95.7% for nuclei and 91.3% for cytoplasm [2], combination of Otsu and Niblack binarization algorithms [3], using Otsu's thresholding methods to segment nucleus and cytoplasm [4], a method based on Gram-Schmidt orthogonalization is proposed along with a snake algorithm to segment nucleus and cytoplasm of the cells [5], this segmentation method achieved an average accuracy is 93%, the combintaion of Otsu's algorithm and watershed algorithm [6], a method base on Geometric Active Contours algorithm to obtain the shape of WBC nuclei [7], Segmentation of White Blood Cells through Nucleus Mark Watershed Operations and Mean Shift Clustering [8]. The accuracy of these methods can be affected when applied to different types of databases, due to changes in cell color, brightness and morphology in the databases.

In studies reporting classification of leukocytes the accuracy is again the key aim for improvement. Muhammad et al. classified 1030 leukocytes into 5 classes [2]. Features were utilized including texture, statistical, wavelet features and wavelet features obtained in the frequency domain. These features were used to classify leukocytes, based on a support vector machine (SVM) with accuracy of 94.3%.

Khamael Abbas et al. used the SVM algorithm in combination with a classification tree to classify 460 WBC images into 10 classes [7]. Here features included cell nuclei mean, scale, kurtosis, skewness. This algorithm gave accuracies between 94.25% for B lymphocytes and 99.01% for neutrophils, this method gave an average accuracy of 97.23%.

Jianwei Zhao et al. used SVM and Random Forest (RF) to classify WBCs into 5 classes [9]. SVM was used to classify basophils and eosinophils by using the features extracted from cytoplasmic particles. Then a convolution neural network (CNN) was used to extract high-level features of the cells. RF was used to classify lymphocytes, monocytes and neutrophils. This method gave an average accuracy of 85.4%.

Seyed Hamid Rezatofighi et al. used the SVM to classify leukocytes into 5 classes [5]. Features were utilized including nucleus and cytoplasm shape, their texture, and color. This method gave an accuracy of 96%.

Merl James Macawile et al. used a CNN to classify 178 cell images into 5 classes [10] with an accuracy of 96.63%.

Anjali Gautam et al, used the Naive Bayes algorithm to classify WBCs into 5 layers. The simple thresholding technique is used for segmentation of leukocytes by using Otsu thresholding. Only the nucleus region was considered for feature extraction. Thereafter, Naïve Bayes classification technique is used for classification of leukocytes, the classification accuracy is about 80.88% [11].

These recent automatic methods for classification of white blood cells have been highly effective, but they required large databases. The above methods are not suitable for small databases, which were collected in Vietnam. Therefore, we propose the Multistage SVM method to improve the above issues.

1.2 Aims

In this paper, the authors proposed a method to classify white blood cells using a Multistage Support Vector Machine. The multistage SVM inherits all the advantages of a regular linear SVM. At each stage of the method, the original database is classified into subclasses. Advantages of the proposed method is that parameters at each stage can be adjusted appropriately to minimize overfitting and achieve the highest accuracy for the available database size.

2. Theoretical basis

2.1 Overview of leukocytes and leukemia diseases

The immune system is a complex system of cells, tissues and organs that work together to protect our bodies from the diseases that cause bacteria, parasites and viruses. Leukocytes perform important roles in the immune system and are classified into 5 major cell types including

neutrophils (50-70%), lymphocytes (25-30%), monocytes (3-9%), eosinophils (0-5%), basophils (0-1%) [10].

In addition, numerous abnormalities in numbers and malignant cells are observed in some blood diseases [12]. Acute leukemia is a group of malignant blood disorders, characterized by blast cells, originating in the bone marrow. Chronic lymphocytic leukemia is characterized by an increasing number of small size mature lymphocytes in peripheral blood, bone marrow and lymph nodes. Plasma cell leukemia is characterized by the appearance of plasma cell lines with a prevalence greater than 20%. Monocytic leukemia is characterized by an increase in the number of monocytes in peripheral blood and in bone marrow.

Diseases are diagnosed based on an increase in the number of WBCs, and the change in shape, color during the development of WBCs into abnormal leukocytes. Each type of WBCs has different characteristics based on the factors of shape, size, color of the nucleus and the cytoplasm. The index extracted from these characteristics is the input of the SVM algorithm used in many articles given in the introduction.

2.2 Multistage support vector machine algorithm

The method proposed in this section extends the original SVM method to a multistage structure. During construction of a SVM model, a subset of samples, known as support vectors, is selected automatically and the discriminant hyperplane with maximum margin is generated The difficult samples will be close to the discriminant hyperplane and the easy-to-classify ones will be further away from it. Based on this, the multistage SVM method bisects the dataset into accepted/rejected subsets with samples "easy-to classify" in the accepted subset and samples "difficult-to-classify" in the rejected subset. The rejected subset is then forwarded to the next stage for further processing, and a second SVM model is trained on the accepted subset. The process resembles a data filtering framework in terms of distinguishing easy and difficult samples and processes them differently. The overall model consists of a chain of successive stages, with two linear SVM models at every stage but the last stage [13].

In order to present the method further, the rules used to stop the iterative process, followed by illustrations of the overall training and testing procedure.

2.2.1 Termination Rules

As the process proceeds, the set of samples that are easy to classify in the underlying model-structure are picked out, leaving only the difficult to classify subset for further processing. As more stages are created, it will be more and more difficult to generate an effective partitioning because the samples left are inconsistent in nature and act like noise. On the other hand, since the partitioning is tail recursive, there will be fewer and fewer samples left, which will increase the risk of overfitting for later stages.

Based on the analysis in [13], two termination rules proposed including:

Rule 1: If the number of the accepted samples is too small, then stop.

Since a second SVM model is trained on the accepted subset, it is important that the size of the accepted subset is sufficiently large. For example, the size should not be smaller than the number of explanatory variables in the model.

Rule 2: Terminate at the earliest stage which maximizes the prediction accuracy.

Cross validation can be used to evaluate the prediction accuracy, and to find the number of stages which maximizes the prediction accuracy. If there are no unique maxima, an earlier stage is preferred to later stages.

2.2.2 Training and testing procedures

The training procedure for the multistage SVM method is described below:

Step 1: Set k: =1 for the first stage. Let S (0) represents the entire data set, and S (k-1) represent the subset forwarded from the previous stage, in other words, the rejected subset at the $k-1^{th}$ stage.

Step 2: Train an initial linear SVM model on the data set S(k-1) for the k^{th} stage. Divide S(k-1) into two parts based on the ± 1 SVM margin. The points falling outside this region are the accepted subset, and the points inside the rejection region are the rejected subset. Then, a second linear SVM model is trained on the accepted subset and saved as the appropriate model for that stage.

Step3: Check the termination conditions. If either of the two conditions is/are satisfied, then stop; otherwise, let k: = k+1, S (k) be the rejected subset of samples and go to step 2.

During testing, we first check to see whether the testing sample is accepted or rejected using the first SVM model at the current stage. If the response value falls inside the rejection region, then the testing sample is rejected; otherwise, it is accepted. The process proceeds, until the testing sample is accepted by a stage (the first stage that accepts it), and the final classification result is calculated by the second SVM model of that stage.

The testing procedure of the multistage SVM algorithm is described below:

Step 1: Set k:=1 for the first stage, and let $f_k(x)$ represent the first linear SVM model, and $g_k(x)$ represent the second linear SVM model for the k^{th} stage.

Step 2: Calculate the value of $f_k(x)$ for the testing sample; if the value is outside the rejection region, then interpret $g_k(x)$ as the final classification label and stop. If the value falls within the rejection region, then let k: = k+1, and do step 2 again.

The training and testing procedures are visualized in Figure 1.

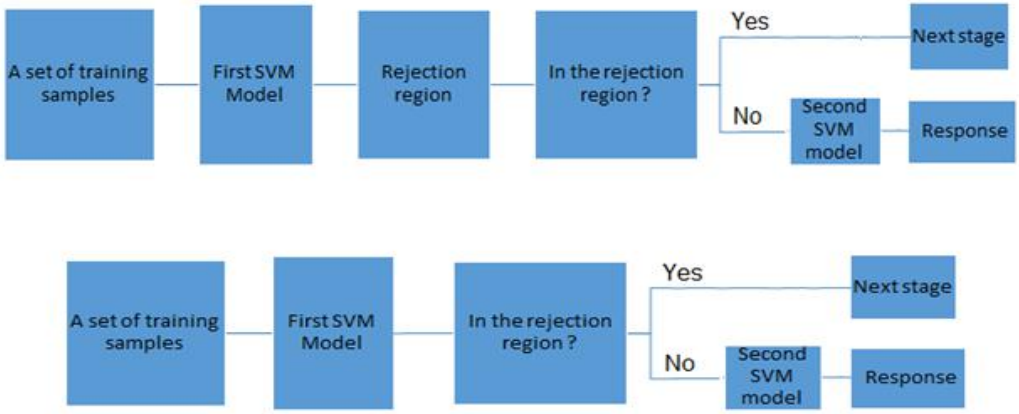

Figure 1: Internal structure of multistage SVM

3. Methodology

Block diagram of implementing the proposed method is shown in Figure 2.

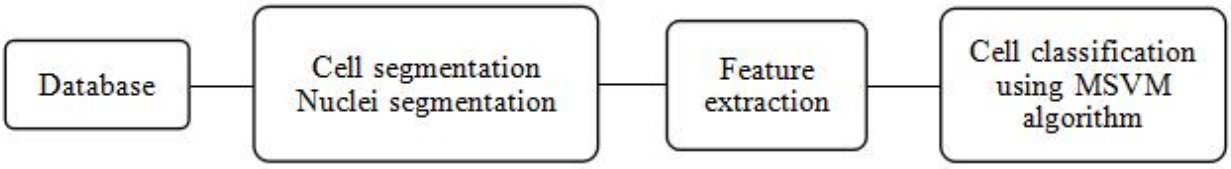

Figure 2: Block diagram of implementing the proposed method

3.1 Cell segmentation

The original images are displayed in RGB color space which is designed to display images in electronic systems. However, the analysis of separate color channels leads to the complexity of the cell segmentation process. Therefore, the original image needs to be transferred to a different color space, which is more convenient for research. The color space system L*a*b is built based on the ability to perceive the color of the human eye. All colors that a normal human eye can see are described by the values L*, a*, b*. In this study, K-means clustering method with the using of two indexes a and b of color space L*a*b is utilized to segment blood cells, this method shows the effectiveness in many researches that has been mentioned above

.

3.2 Cell nuclei segmentation

Geometric Active Contours algorithm (GACs) algorithm is used to cell nuclei segmentation. The implementation process and theoretical basis are given in references [14], [15].

The results of nucleus and cytoplasm segmentation are converted to grayscale as shown in Figure 3.

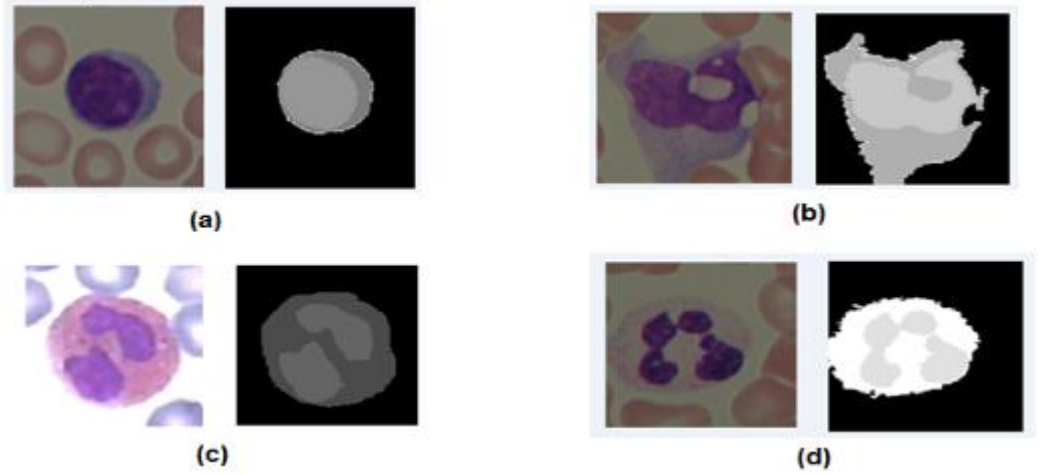

Figure 3: (a) Lymphocyte, (b) Monocyte (c) Eosinophil (d) Neutrophils

The authors used existing methods, mentioned in the references to segment the white blood cell and the nuclear region of white blood cell. The analyses are carried out to evaluate the performance of the white blood cell pixels detection by comparing with the pixels of manually marked white blood cell.

Evaluation of segmentation technique is based on three metrics including precision, recall, and F-measure which can be computed as follow:

Measures	Formula
Precision	$\frac{TP}{TP + FN}$
Recall	$\frac{TP}{TP + FN}$
F-measure	$2*\frac{Precision* Recall}{Recall+ Precision}$

Where

-True positive (TP) is the number of WBC pixels correctly identified.

-False positive (FP) is the number of non-WBC cells pixels that are marked as WBC pixels.

-False negative (FN) is the number of WBC pixels incorrectly labelled as non-WBC

-F-Measure is the weighted harmonic mean of precision and recall.

Table 1: Results of Cell segmentation use K-means clustering method

Type of WBC	Precision	Recall	F-measure
Lymphocyte	0.96	0.99	0.97
Monocyte	0.94	0.97	0.95
Neutrophil	0.93	0.95	0.94
Eosinophil	0.90	0.88	0.89

Table 2: Results of Cell nuclei segmentation use Geometric active contours method

Type of WBC	Precision	Recall	F-measure
Lymphocyte	0.99	0.95	0.94
Monocyte	0.99	0.9	0.93
Neutrophil	0.99	0.85	0.91
Eosinophil	0.98	0.81	0.88

From the two tables of segmentation results, it is shown that the process of cell segmentation and nuclear cell segmentation has achieved high results, indicating efficient segmentation results closer to manual segmentation. Achieving good results in the segmentation process will help increase the accuracy of the features extracted from the cell and the nucleus of white blood cells. There by increasing the accuracy of white blood cell classification results. The results of the segmentation also depend on the manual segmentation process and the quality of the white blood cells image after being stained.

3.3 Feature extraction

There are many cell characteristics used in references, namely shape, size, color, statistical indicators of nucleus and cytoplasm. In this study, cell features used include: cytoplasm color, cell size, nucleus/cytoplasmic ratio, roundness and firmness of cell nucleus.

3.4 White blood cell classification using Multistage SVM algorithm

In this study, the authors used the MSVM algorithm to classify 389 white blood cells into 4 classes: eosinophils, neutrophils, monocytes, lymphocytes (small lymphocytes, large lymphocytes, lymphoblast), basophils are not classified due to the low number of samples in the database. The data is divided into 2 subsets with a ratio of 3: 1 for training and testing.

At each stage, different features are used to classify, the stages of the algorithm are shown as follows:

Stage 1: Classify eosinophils with the remaining cells based on the features of the average index and standard deviation of cytoplasm in RGB color space.

Stage 2: Classify neutrophils with remaining cells based on the features of the nucleus in terms of roundness and firmness

Stage 3: Classify monocytes with lymphocytes based on the features of the nucleus in terms of roundness, firmness, cell area, ratio of nucleus to cytoplasm.

Model of Multistage SVM Classification is shown in Figure 4

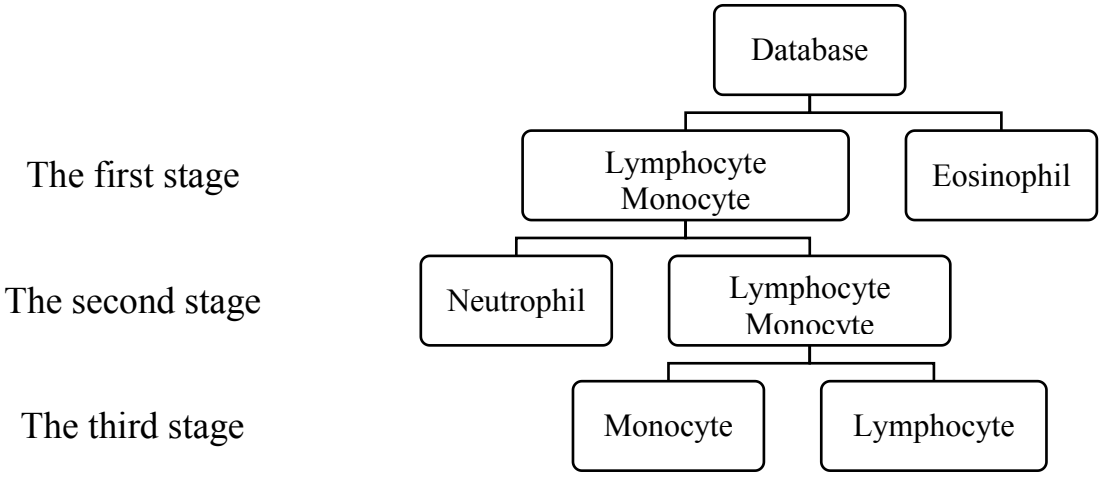

Figure 4: Model of Multistage SVM Classification

3.5 Database

The database used in this paper includes 389 white blood cells collected from the database ALL_IDB (Acute Lymphoblastic Leukemia images database) and LISC (Leukocyte Images for Segmentation and Classification) including 237 cells. lymphocytes, 46 monocytes, 67 neutrophils, eosinophils.

4. Results and discussion

The accuracy of each stage and the combined model of the three stage after practice is given in Table 3:

Table 3: Classification results of the stages and the 3-stage model

	Training				Testing			
Stages	Total	Miss	Correct	Accuracy (%)	Total	Miss	Correct	Accuracy (%)
1	295	0	295	100	94	0	94	100%
2	265	14	251	94.7	85	1	84	98.8
3	215	10	205	95.3	68	3	65	95.6
3-stage model	295	23	272	92.2	94	8	86	91.4

At each stage of the algorithm, a model is utilized to evaluate subset, and test set of each stage. Then, to form a multi-stage classification model, a combination of 3 models of the stages is built to evaluate the entire testing database. The accuracy of the classification model at each 1, 2, 3 stage is 100%, 98.8% and 95.6%. The accuracy of the combination model is 91.4%.

The results of the proposed method are compared with SVM method combined with Random forest method [9], Naïve Bayes method [11] and are shown in Table 4.

Table 4: Comparing results of proposed method with other methods

Cell types / Methods	Lymphocyte	Monocyte	Neutrophil	Eosinophil	Basophil	Average accuracy
Combination of SVM and Random forest [9]	74,8%	85,3%	97,1%	70%	100%	81,8%
Naïve Bayes [11]	100%	90%	55%	87,5%	60%	79%
Proposed method	92%	91%	89%	100%	—	93%

The author used this combination model to classify for each cell type, the accuracy of method when classifying lymphocytes, monocytes, neutrophils, Eosinophils is 92%, 91%, 89%, 100%. The average accuracy is 93%.

The results of the proposed method with the results of Naïve Bayes method [11] and SVM method combined with Random forest [9] are compared because two methods both classify leukocytes into 5 main classes (lymphocyte, monocyte, neutrophil, eosinophil and basophil). In addition, the database of reference [9] also used the database ALL_IDB as input image for classification process, and authors in reference [11] also extracted features from the

shape and size of white blood cells. The selection of features is mentioned because it plays an important role in the classification of images to achieve high results. From Table 2, it is indicated that the results of the proposed method for eosinophil and monocyte classification obtained the highest accuracy among three methods. Although the accuracy of neutrophil and lymphocyte classification of the proposed method is lower than the accuracy of this leukocyte classification of the other two methods (8%), however the results of classifying leukocytes into 4 different classes of the proposed method obtained the highest accuracy (93%). This shows the high efficiency of the proposed method for leukocyte classification and may be of most use to aid clinicians-for example they may use this automated system to quickly classify leukocytes then focus their time on results indicating blood disorders or unknown cell morphologies.

5. Conclusion

In this paper, Multistage SVM algorithm is utilized to classify cells based on the features of shape, size, color, area ratio of nucleus and cytoplasm. The authors used two databases to perform training and testing and from the results showing the effectiveness of the method with mixed databases, the time of training process only takes a few minutes, average accuracy is high 93%. A number of incorrectly segmented cells and the influence of the cell growth stages in lymphocytes causing differences in shape, size of cells lead to incorrectly classified cells.

We show that the features used in the classification process are suitable for classifying white blood cells, and these are clinician informed as the features are based on the characteristics of the cells that humans use to classify.

The authors plan to develop the proposed method in combination with other methods and will perform classification of blood cells in a database collected from the National Institute of Hematology and Blood Transfusion in VietNam in the near future.

References

1 Oleg, R.; Anuradha, R.; Thomas, B.; Martin, F.; Stefan, H.; Claus, K.; Michael, B,; Ute, N.; Juergen, P. Leukocyte subtypes classification by means of image processing, ***Proceedings of the Federated Conference on Computer Science and Information Systems***, 2016, 8, 309 – 316 [DOI: 10.15439/2016F80]

2 Muhammad, S.; Siraj, Kh; Zahoor, J.; Khan, M.; Hyeonjoon, M.; Jin Tae, K.; Seungmin Rho; Sung Wook, B.; Irfan, M. Microscopic Blood Smear: A Resource-Aware Healthcare Service in Smart Cities, ***IEEE***, 2016, 1-15 [DOI: 10.1109/ACCESS.2016.2636218]

3 Mehdi, H.; Adam, K.; Thomas, F. Comparative study of shape, intensity and texture features and support vector machine for white blood cell classification, ***Journal of Theoretical and Applied Computer Science***, 2013, 7(1), 20-35.

4 Subham, M.; Lalit Mohan, S.; Nikhil, V. Counting and Classification of White Blood Cell using Artificial Neural Network (ANN), ***IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems***, 2016, 1-5 [DOI: 10.1109/ICPEICES.2016.7853644]

5 Rezatofighi SH, Soltanian-Zadeh H. Automatic recognition of five types of white blood cells in peripheral blood. ***Comput Med Imaging Graph*** 2011; 35(4): 333-343 DOI: 10.1016/j.compmedimag.2011.01.003

6 Simge Çelebi and Mert Burkay Çöteli. Red and white blood cell classification using Artificial Neural Networks, ***AIMS Bioengineering***, 2018, 5(3), 179-191, [DOI: 10.3934/bioeng.2018.3.179]

7 Khamael AL-Dulaimi; Kien, N.; Jasmine, B.; Vinod Chandran; Inmaculada, Tomeo-R. Classification of White Blood Cells Using L-Moments ***Invariant Features of Nuclei Shape***, 2018 [DOI: 10.1109/IVCNZ.2018.8634678]

8 Liu Z, Liu J, Xiao X, Yuan H, Li X, Chang J, Zheng C. Segmentation of White Blood Cells through Nucleus Mark Watershed Operations and Mean Shift Clustering. ***Sensors (Basel)*** 2015; 15(9): 22561-22586 [PMID: 26370995 PMCID: PMC4610533 DOI: 10.3390/s150922561]

9 Zhao J, Zhang M, Zhou Z, Chu J, Cao F. Automatic detection and classification of leukocytes using convolutional neural networks. ***Med Biol Eng Comput*** 2017; 55(8): 1287-1301 DOI: 10.1007/s11517-016-1590-x

10 Merl James, M.; Vonn Vincent, Q.; Alejandro Ballado Jr.; Jennifer Dela Cruz; Meo Vincent, C. White Blood Cell Classification and Counting Using Convolutional Neural Network, ***The 3rd International Conference on Control and Robotics Engineering***, 2018, 259-263 [DOI: 10.1109/ICCRE.2018.8376476]

11 Anjali, G.; Priyanka S.; Balasubramania, R.; Harvendra Bhadauria. Automatic Classification of Leukocytes using Morphological Features and Naïve Bayes Classifier, ***IEEE Region 10 Conference (TENCON)-Proceedings of the International Conference***, 2016, 1023-1027 [DOI: 10.1109/TENCON.2016.7848161]

12 Harald Theml, Heinz Diem, Torsten Haferlach. Color Atlas of Hematology Practical Microscopic and Clinical Diagnosis. ***Theml Stuttgart: New York***, 2004; pp 1-189

13 https://docs.lib.purdue.edu/ecetr/

14 Khamael AL-Dulaimi; Inmaculada Tomeo-Reyes; Jasmine, B.; Vinod Chandran. White blood cell nuclei segmentation using level set methods and geometric active contours, ***International Conference on Digital Image Computing: Techniques and Applications (DICTA)***, 2016, 1-7 [DOI: 10.1109/DICTA.2016.7797097]

15 Khamael AL-Dulaimi; Inmaculada Tomeo-Reyes; Jasmine, B.; Vinod Chandran. Automatic segmentation of hep-2 cell fluorescence microscope images using level set method via geometric active contours, ***23rd International Conference on Pattern Recognition (ICPR)***, 2016, pp 81-83 [DOI: 10.1109/ICPR.2016.7899612]

