Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2019) 11(1), 141-152.

10.29137/umagd.498372

Cilt/Volume:11 Sayı/Issue:1 Ocak/January 2019

Araştırma Makalesi / Research Article

Classifying White Blood Cells Using Machine Learning Algorithms

Abdullah Elen *[1], M. Kamil Turan [2]

[1]*Karabük University, Dept. of Comp. Tech., Vocational School of T.O.B.B. Tech. Sci., 78050 Karabük, TURKEY*
[2]*Department of Medical Biology, Faculty of Medicine, Karabük University, 78050 Karabük, TURKEY*

***Başvuru/Received**: 17/10/2018* ***Kabul/Accepted**: 21/12/2018* ***Son Versiyon/Final Version**: 31/01/2019*

Abstract

Blood and its components have an important place in human life and are the best indicator tool in determining many pathological conditions. In particular, the classification of white blood cells is of great importance for the diagnosis of hematological diseases. In this study, 350 microscopic blood smear images were tested with 6 different machine learning algorithms for the classification of white blood cells and their performances were compared. 35 different geometric and statistical (texture) features have been extracted from blood images for training and test parameters of machine learning algorithms. According to the results, the Multinomial Logistic Regression (MLR) algorithm performed better than the other methods with an average 95% test success. The MLR can be used for automatic classification of white blood cells. It can be used especially as a source for diagnosis of diseases for hematologists and internal medicine specialists.

Key Words

"WBC classification, leukocytes, blood cells, machine learning."

***Corresponding Author:** aelen@karabuk.edu.tr*

1. INTRODUCTION

Blood is a structure consisting of plasma and blood cells in the circulatory system of the heart and veins, which we call the cardiovascular system in the body. Microscopic analysis of peripheral blood smear results in hematology is a costly and time-consuming process (Krzyzak et al. 2011, Li et al. 2014, Maji et al. 2015). White blood cells (*leukocyte, WBC*) often lead to misidentification and classification because they are not inherently stable (Pandit, A., Kolhar, S. & Patil, P. 2015, Rawat et al. 2015). For this reason, it is highly probable that blood tests with traditional methods will encounter these problems. At the same time, due to the statistical bias and inconsistencies (Sonar, S. C. & Bhagat, K. S. 2015) of the analyst (hematologist), both the subjective evaluation of results and slow progress of the process (Xiong et al. 2010, Tomari et al. 2014, Venkatalakshmi, B. & Thilagavathi, K. 2013). For these reasons, the development and use of computer-based systems instead of traditional methods will greatly contribute to the acceleration of the analysis process and more accurate results (Maji, P., Mandal, A., Ganguly, M. & Saha, S. 2015, Pandit, A., Kolhar, S. & Patil, P. 2015).

Researchers are increasingly interested in the development of algorithms for automated analysis of medical images such as microscopic blood smear. Researchers working on image processing, computer vision, artificial neural networks, machine learning algorithms etc. techniques for blood cell analysis. Some of the studies in the literature are as follows; Sanei and Lee (Sanei, S. & Lee, T. K. M. 2003), in their study, aimed to improve the work of Turk and Pentland in the selection of eigenvector from monochrome images. Accordingly, instead of monochrome image, they used three components in color image. With the Bayesian classifier, they classified the Eigen cells, not the physical or geometric properties of the image. They used density and color information as parameters in the decision-making process. First, they rescanned the input images, then segmented and rotated, and finally identified 3 vectors representing the intensity and color information. Sarrafzadeh et al. (Sarrafzadeh, O., Rabbani, H., Talebi, A. & Yousefi-Banaem, H. 2014) used 149 leukocyte images of 10 patients. As the classifier, 6 geometric properties, 6 color properties, 6 statistical properties and 7 moment invariance (invariant) were used as parameters for SVM. The classifier reported that they achieved over 93% success. Leukocyte boundaries in images are manually determined to reduce the effects of segmentation errors. The nucleus and cytoplasm of leukocytes were automatically separated by the Fuzzy C-means clustering method. Then appropriate properties are extracted from the nucleus, the cytoplasm and the cell. These properties are classified by SVM. Ko et al. (Ko, B. C., Gim, J. W. & Nam, J. Y. 2011) used half of the 240 blood smear images for train and the other half for testing. They preferred the Random Forest method in the classification process and claimed that they achieved higher success than Multi-layer SVM. In their previous studies, GVF recommended leukocyte segmentation with the Snake algorithm. They used the shape, color and texture properties of the image as the classification parameter. After the feature extraction, they normalized each feature vector from 0 to 1 with Gaussian normalization. Ramoser et al. (Ramoser, H., Laurain, V., Bischof, H. & Ecker, R. 2005) used SVM for automatic leukocyte grading. In the evaluation made in 1166 imagery group consisting of 13 different classes, 95% correct segmentation and 75-99% accurate classification were made. Theera-Umpon and Dhompongsa (Theera-Umpon, N. & Dhompongsa, S. 2007) have investigated whether it is enough to classify leukocytes only with information from the leukocyte nucleus. In order to prevent segmentation errors in the experiments, they manually removed the cell nuclei. They used Bayes classifiers and artificial neural networks for classification. They reported that the information obtained from the cell nucleus was enough according to the 77% success of their classification. Adjouadi et al. (Adjouadi, M., Zong, N. & Ayala, M. 2005), in their study, proposed a method for the type identification of white blood cells in flow cytometry. They analyzed the behavior of parametric datasets in a multidimensional range using Support Vector Machines (SVM). Rodrigues et al. (Rodrigues, P., Ferreira, M. & Monteiro, J. 2008) used an artificial neural network consisting of two stages in the classification of white blood cells. In the first stage, they applied a pre-classification by applying back propagation algorithm (BPNN), and in the second step, they presented a hybrid model using the support vector machine (SVM) and Puls-Coupled neural network (PCNN) to reduce the detected problems. Thus, they aimed to minimize the negative aspects. Joshi et al. (Joshi, M. D., Karode, A. H. & Suralkar, S. R. 2013) proposed the Otsu's automatic thresholding algorithm for segmentation of blood cells and the image enhancement and arithmetic method for leukocyte segmentation. K-NN classifier was used to classify blast cells from normal lymphocyte cells. They obtained a 93% accuracy rate according to the test results. Tantikitti et al. (Tantikitti, S., Tumswadi, S. & Premchaiswadi, W. 2015), in their study, used image processing techniques such as color transformation, image fragmentation, edge detection feature extraction, and white blood cell classification. They classified the dengue virus infections of patients with decision tree method. According to the results obtained, they reported that 167 cell images were successful in leukocyte classification with 92.2% and 264 blood cell images with 72.3% accuracy in dengue classification. Saraswat and Arya (Saraswat, M. & Arya, K. V. 2014) used Random Forests to classify leukocyte cells as mono-nuclear and polymorph-nuclear cells from blood smear images obtained with 40X magnification.

In the classification of white blood cells, color, texture and geometric properties of the images were used as input parameters of artificial intelligence-based algorithms. Some of the studies in the literature are as follows; Hiremath et al. (Hiremath, P. S., Bannigidad, P. & Geeta, S. 2010) for lymphocytes, monocytes and neutrophil cells only; histogram equalization, edge extraction and threshold-based automatic segmentation. Geometric properties of the images were used for the classification process and 100 different blood smear images were used in the experiments. Habibzadeh et al. (Habibzadeh, M., Krzyzak, A. & Fevens, T. 2013) aimed to classify and count leukocytes according to 5 different categories by using the shape, density and texture properties of microscopic blood images. The wavelet properties obtained by the Dual-Tree Complex Wavelet Transform (DT-CWT) method for the classification process were used as parameters of the SVM classifier. Ramesh et al. (Ramesh, N., Dangott, B., Salama, M. E. & Tasdizen, T. 2012) proposed a simple classification method using color information and morphological features. As the first step in a two-stage classification process, they have broadly classified leukocyte cell nuclei and leukocyte boundaries. In the second

step, they used the properties obtained from leukocyte cytoplasm and nucleus by Linear Discriminant Analysis method. Ferri et al. (Ferri, M., Lombardini, S. & Pallotti, C. 1994) used dimensional functions using the morphological features of images for the automatic classification of leukocyte cells. Bikhet et al. (Bikhet, S. F., Darwish, A. M., Tolba, H. A. & Shaheen, S. I. 2000) tested the algorithm according to the selected properties and obtained an accurate classification rate of more than 90%. Selected features; cell area, nucleus area, cytoplasm area, the ratio of nucleus to cell area, average color of cytoplasm, the ratio of nucleus area to its environment, cell circularity, nucleus circularity and number of nucleus. Su et al. (Su, M., Cheng, C. & Wang, P. 2014), according to the study, five different types of leukocyte cells in the HSI color space to distinguish the distinctive properties of leukocytes. The elliptic areas in this region were leukocyte nuclei and cytoplasm, and then they aimed to segment by morphological processes. From these image segments, geometric features, color characteristics and LDP (*Local Directional Pattern*) based texture properties were removed and trained in three different neural networks. In the classification tests, they used 450 leukocyte images and found the highest accurate identification rate as 99.11%.

This study focuses on the classification of segmented leukocyte cells obtained from microscopic blood smear images. Its performance was compared by using machine learning algorithms for classification operations. The statistical and geometrical feature information obtained from the images were used as input parameters of the algorithms. The rest of the paper is organized as follows; In the second part, the morphological features of white blood cells, the geometric and RGB color space of the blood image based on the extraction of statistical features and 6 different machine learning algorithms used in the study is mentioned. In the third chapter, the test results of the 350 blood smear images as a dataset and the machine learning algorithms used in the experiment are mentioned comparatively. In the last section, the whole process is evaluated in general; According to the results obtained, the most suitable machine learning algorithm and some other suggestions for the classification of leukocyte cells are mentioned.

2. MATERIALS AND METHODS

In our previous study (Elen, A. & Turan, M. K. 2018), microscopic blood smear images were segmented, and the blood cells were divided into three main groups: erythrocytes, platelets and leukocytes. We used the leukocyte images to be divided into five different classes by machine learning algorithms. Statistical and geometric features of WBC images were obtained for the input parameters of machine learning algorithms.

2.1. White Blood Cells

White blood cells (WBC), also called leukocytes, are produced in the bone marrow. Leukocyte cells are composed of nuclei and cytoplasm. They are divided into five groups: basophil, eosinophil, lymphocyte, monocyte and neutrophil. Leukocytes, which protect the body against infectious diseases and foreign substances, constitute an important part of the immune system. 4×10^9-11×10^9 units in one liter of a healthy adult human. That is, a drop in the blood is about 7000 to 25000. Figure 1 shows the average number of white blood cells in a healthy adult. Neutrophils are the most common leukocytes in human blood. The kernels consist of 3-5 lobes. Polymorphonuclear constitute 99% of the cells while the polymorphonuclear cells account for about 70% of the total leukocyte count. Eosinophils grab onto lots of eosin dye (a type of acid red dye) when they're stained, making their large granules a red color. Their lifespan is 1-2 weeks and constitute 2-3% of all leukocytes. They have an average diameter of 10-12 µm and their nuclei are two lobes.

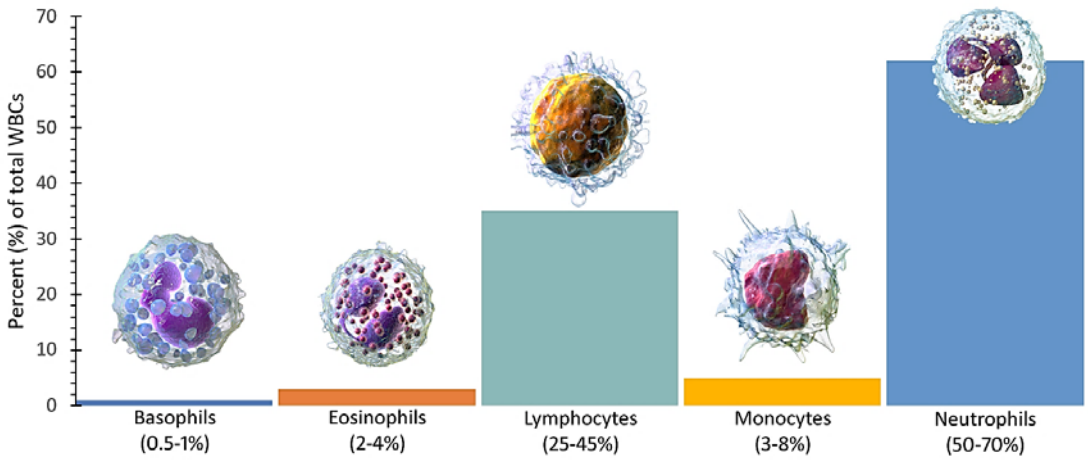

Fig. 1. Average values for a normal adult white blood cell count.

Another group of leukocytes called as basophil, its granules grab onto lots of basic dyes and have deep blue-purple color. Its nucleus is irregular and consists of two lobes that cannot be distinguished. It is also the least number of leukocytes. Monocytes are the largest cells of peripheral blood (15–22 µm). Folds can be seen in the nucleus, which can be of different shapes (*round, lobular, kidney, bean or horseshoe*). Lymphocytes are cells that can divide and give new lymphocytes. When they encounter immunogenic (*antigenic*) stimulation; morphological transformation, differentiation and multiplication. It is the most common type of leukocytes

in the blood after neutrophils. In this study, 350 different WBC images were used as the dataset. Figure 2 shows sample leukocyte cell images used in the classification process.

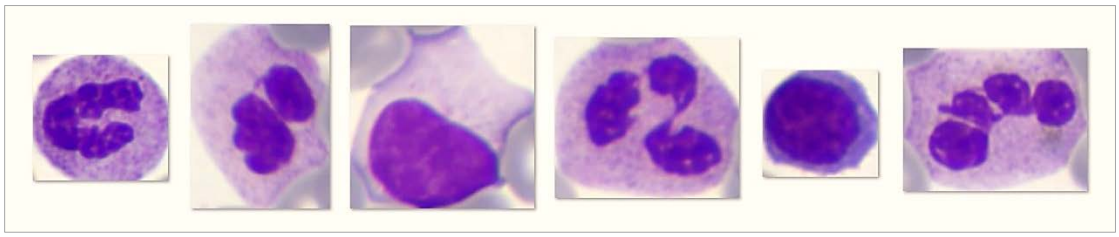

Fig. 2. Sample images of leukocyte cells used in the study.

2.2. Feature Extraction

Feature extraction in image processing is a method of converting large amounts of unnecessary data into a reduced data display. The process of converting input data into a property dataset is called feature extraction (Krishnan, A. & Sreekumar, K. 2014, Avuçlu, E., & Başçiftçi, F. 2018). Feature extraction methods analyze objects and images to extract the most distinctive features representing various object classes. Property vectors are used as input parameters to classifiers assigned to the class to which they are represented. The purpose of feature extraction is to reduce the original data by scaling certain properties or properties that distinguish an input set from another set. Feature extraction is an important process in the automatic classification of white blood cells (Rawat, J., Bhadauria, H. S., Singh, A. & Virmani, J. 2015) and the selected properties affect the performance of the classifiers. The accuracy of the classification depends on the number of features, and feature properties.

An important part of the studies on microscopic images for the classification of leukocyte cells has been based on geometric and tissue-based properties (Osowski, S., Siroic, R., Markiewicz, T. & Siwek, K. 2009). The geometric properties used to distinguish cells include shape and size of nucleus, shape and size of the white blood cell, number of nucleus lobes, cell circularity, and nucleus rectangularity (Rosin, P. L. 2003). Tissue is the specific granule and chromatin-induced properties in the nucleus. The texture feature (Tuceryan, M. & Jain, A. K. 1998) includes statistical information such as mean, standard deviation, skewness, kurtosis and entropy of brightness. In this study, 35 different statistical and geometric characteristics were used to classify leukocyte cells. In the previous study, we used the reference polygons of the WBC cells (cytoplasm and nucleus) we segmented. Figure 3 shows the segmentation steps of WBC cells.

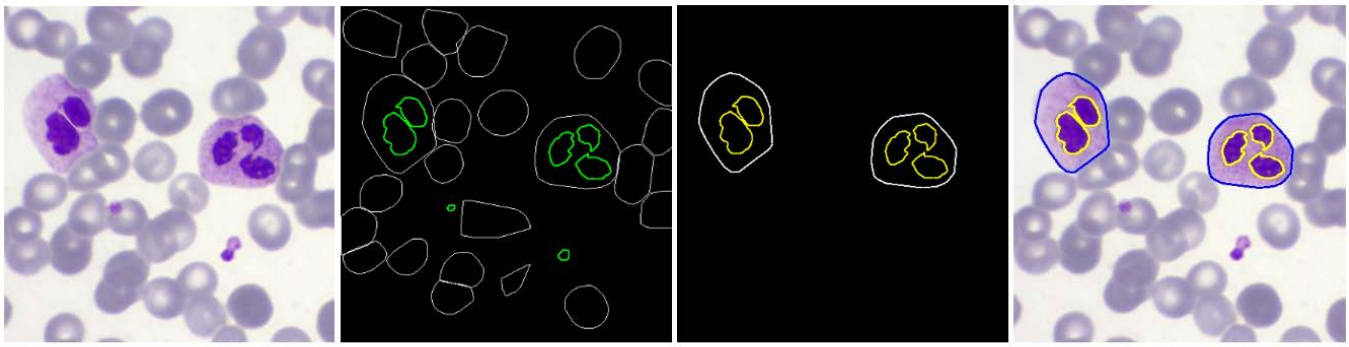

Fig. 3. Segmentation of WBCs; a) input image, b) segmentation of blood cells, c) extraction of WBCs, d) result image.

The geometric and statistical properties of the machine learning algorithms used as a parameter for the training and testing of leukocyte cells are given in Figure 4. In order to extract the statistical features, the white blood cell images were calculated using the histogram values for each color channel according to the RGB color space. Accordingly, for nucleus and leukocyte features; A feature vector ($F_S = \{p1, p2, \dots, p21\}$) was created by using the 21 parameters including kurtosis, mean and standard deviation values as well as energy value for the nucleus. In order to extract the geometric features, WBC images were first converted to grayscale and then, for both leukocyte and nucleus, radius, area, perimeter, ratio of leukocyte perimeter to nucleus perimeter (P_L/P_N) and ratio of leukocyte area to nucleus area (A_L/A_N) parameters were used. In addition, a feature vector ($F_G = \{p1, p2, \dots, p14\}$) was created using 14 parameters, including the number of lobes in the leukocyte nucleus, circularity for the nucleus, fullness, and compactness values. Thus, both geometric and statistical feature vectors were combined to form a feature vector ($FV = \{F_S, F_G\}$) consisting of a total of 35 parameters for the classification algorithms.

	Leukocyte	Nucleus
Radius	+	+
Area	+	+
Perimeter	+	+
Circularity	—	+
Fullness	—	+
Compactness	—	+
Number of lobes	+	—
Perimeter: P_L / P_N	*Ratio of leukocyte to nucleus for perimeter and area.*	
Area: A_L / A_N		

a)

		R	G	B
Leukocyte	Std. Deviation	+	+	+
	Mean	+	+	+
	Kurtosis	+	+	+
	Energy	—	—	—
Nucleus	Std. Deviation	+	+	+
	Mean	+	+	+
	Kurtosis	+	+	+
	Energy	+	+	+

b)

Fig. 4. Features used in the classification of WBCs; a) geometric features, b) statistical features.

2.3. Machine Learning Algorithms

Machine Learning is the method paradigm that makes inferences from the available data using mathematical and statistical methods and makes predictions about the unknown with these inferences. Machine learning is one of the fastest growing areas of computer science with a wide range of applications. Some academic research in the past have shown that after a certain stage, the machines must be learned the data. As a result of this, researchers carried out their studies in order to approach various problems by using various symbolic methods (Sarle, W. S. 1994). A significant number of these approaches have ability to estimation, prediction and classification. In this section, the properties of machine learning algorithms that used in this study for classification of WBCs are mentioned.

2.3.1. Decision Tree Classifier

The Decision Tree is a consulted machine learning algorithm that can classify data by continually dividing the dataset according to a certain criterion. A decision tree structure consists of roots, nodes, branches and leaves. The bottom part of the tree structure and the upper part of the leaves are called roots. Each feature in the dataset represents nodes. The link between the nodes is called the branch. It is very important to decide which node to start partitioning in decision trees. If the appropriate node does not start, the number of nodes and leaves in the tree will be very high. Many decision tree learning algorithms are available in the literature. In this study, C4.5 algorithm was preferred. The entropy of the class attribute is calculated first for this operation, as shown in Equation 1.

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 (p_i) \tag{1}$$

In the next step, as shown in Equation 2 and Equation 3, the class-dependent entropies of the feature vectors (X) to the class (S) are calculated.

$$H(X_k) = -\sum_{i=1}^{n} \frac{|S_i|}{|X_k|} \log \frac{|S_i|}{|X_k|} \tag{2}$$

$$H(X,S) = -\sum_{k=1}^{n} \frac{|X_k|}{|X|} H(X_k) \tag{3}$$

Finally, the entropy of all feature vectors is calculated by the entropy of the class attribute and the gain metric for each property is calculated. Knowledge gain, Gini index and Towing rule are commonly used methods. In this study, Knowledge Gain method was used as decision criteria (Equation 4).

$$IG(X,S) = H(S) - H(X,S) \tag{4}$$

2.3.2. Random Forest

The Random Forest algorithm was developed by Breiman in 2001 (Breiman, L. 2001). In this method, instead of producing a single decision tree, it combines the decisions of many multivariate trees, each trained in different sets of training. As a result, it is an algorithm that achieves high levels of success in solving classification problems. In the Random Forest algorithm, the determination of branching criteria and the selection of a suitable pruning method as in the other decision tree methods are an important issue. Gain ratio and Gini index are the most commonly used gain measurement techniques in determining the branching criteria. The operation of this algorithm is based on two different parameters: the number of trees to be developed and the number of samples

used for each node. In the classification process, primarily the user-defined tree is created. When a new sample is to be classified, it is treated by the decision tree and the class of the new sample is determined according to the highest rate obtained from these trees (Pal, M. 2005).

2.3.3. k-Nearest Neighbors (k-NN)

The k-NN algorithm was proposed by Cover and Hart in 1967 (Cover, T., & Hart, P. 1967). k-NN is one of the most basic pattern recognition and classification methods that classify objects according to the nearest training instances in the attribute space. The aim here is to decide that new sample belongs to which class, according to the k value of the nearest neighbor. To determine the class of a new vector, the closest k samples selected from the training data are selected. Accordingly, the new vector is assigned to it by looking at the classes in which the selected samples belong. A new example has different methods (*Euclid, Manhattan, Minkowski, etc.*) for calculating distances according to classified samples. The most common of these is the Euclidean distance calculation method (Equation 5). Where i and j are two input instances and k is the number of neighbors in Equation 5.

$$d(i,j)=\sqrt{\sum_{p=1}^{k}\left(X_{ip}-X_{jp}\right)^2} \tag{5}$$

2.3.4. Multinomial Logistic Regression (MLR)

Regression analysis is a statistical method used to determine the relationship between two or more variables with cause-effect relationship and to make estimations or predictions on the subject by using this relationship. Logistic regression (LR) is a nonlinear regression model designed for two dependent variables. MLR is used to describe cause-and-effect relationships between dependent variable (Y) and independent variables (X) where the dependent variable contains at least three or more categories (Leech, N. L., Barrett, K. C. & Morgan, G. A. 2004) and the values are obtained by a classifying scale (Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. 2013, Washington, S. P., Karlaftis, M. G. & Mannering, F. 2003, Arı, E., & Yıldız, Z. 2013).

The purpose of this analysis is to estimate the value of categorically dependent variables, so this is an estimate of membership for two or more categories. Accordingly, one of the objectives of the method is to classify and the other to investigate the relationships between dependent and independent variables (Büyüköztürk, Ş., Çokluk Bökeoğlu, Ö. & Şekercioğlu, G. 2010). The LR model is a special form of general linear models obtained for the dependent variables with binomial distribution and is expressed as in Equation 6;

$$\pi(x)=\frac{\alpha+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p}{1+e^{\alpha+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p}} \tag{6}$$

Here, $\pi(x)$ represents the probability of an event being examined, α dependent variable constant, $\beta_1,\beta_2,\ldots,\beta_p$ independent variables regression coefficients, $x_1,x_2,\ldots,x_p$ arguments, p independent variable number and e error term. The MLR model is the extended version of the LR model with two states, as shown in Equation 7.

$$\pi_j(x_i)=e^{\alpha_i+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_p}\Bigg/\left(1+\sum_{j=1}^{k-1}e^{\alpha_i+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_p}\right) \tag{7}$$

Here, $j_1,j_2,\ldots,j_k$ represents k category, n $(i_1,i_2,\ldots,i_n)$ represents the level of possible independent levels.

2.3.5. Naïve Bayes

The Naïve Bayes Classifier is a simple probabilistic classification method based on Bayes' theorem (Thomas Bayes, 1702-1761). In the case of Bayes' theorem, in the case of two random events (X and Y) occurring consecutively, the probability of the occurrence of the second event in the event of one of these two events can be represented by $P(X\cap Y)$. As with Equation 8, the multiplication rule can be written with two different expressions;

$$P(X\cap Y)=P(X|Y)P(Y)=P(Y|X)P(X) \tag{8}$$

The Bayes' theorem describes the relationship between an arbitrary X event due to a random process and conditional probabilities and marginal probabilities for another random Y event (Equation 9).

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{9}$$

The probabilities of the dependent situations likely to occur in any problem are calculated by the Bayes equation given above. In this equation, the $P(X)$ expression represents the probability of the problem input, the probability of the $P(Y)$ statement possible output state, and the $P(Y|X)$ expression represents the probability of Y output states versus the previous X input (Orhan, U. & Adem, K. 2012). In the Naïve Bayes classification technique, it analyzes the relationship between dependent and independent features to create a contingent probability from each relationship. To classify a new instance, an estimate is made by combining the effects of the independent variables on the dependent variable (Krishna, P. R. & De, S. K. 2005).

2.3.6. Support Vector Machine (SVM)

The Support Vector Machine is a machine learning algorithm based on the principle of structural risk minimization and based on convex optimization (Soman, K. P., Loganathan, R. & Ajay, V. 2009). It is mainly designed to solve binary classification problems. The aim here is to obtain a hyperplane that will optimally separate the classes from each other. In the classification, it is usually represented by class labels such as {-1, +1}. The data to be classified can be separated linearly (AND/OR problem) or cannot be separated by a single line (XOR problem). Therefore, SVM is divided into two groups as Linear SVM and Nonlinear SVM depending on the data. As is known, many classification problems in the real world consist of more than two classes. To solve such problems, a multi-class SVM classifier is needed. Multiple classification can be achieved by combining binary classifiers (Jiang, Z. G., Fu, H. G. & Li, L. J. 2005). If it is assumed that the training data consisting of n numbers of samples for training of SVM in a linearly separable class classification problem is $\{x_i, y_i\}, i = 1,2, \dots, n, y_i \in \{-1, +1\}, x_i \in R^d$, then the decision function of the optimal separation plane will be as in Equation 10;

$$y_i = \begin{cases} w \cdot x_i + b \geq +1, & +1 \\ w \cdot x_i + b \leq -1, & -1 \end{cases} \tag{10}$$

Where R^d represents the D-dimensional space of the input patterns (x_i), y_i represents the labels where the inputs are classified as {-1, +1}. w represents the normal value of the multiple plane, b represents the tendency (*bias*) value. In order to determine the optimal separation plane, the boundaries that are parallel to this correction must be determined. That is, support vectors are required. This procedure is expressed as $w \cdot x_i + b = \pm 1$ (Kavzaoğlu, T. & Çölkesen, İ. 2010). As in the classification of medical images, it is not possible to separate the data linearly in many other image processing problems. In this case, it is possible to solve the problem by defining a part of the training data on the other side of the optimal hyperplane by defining a positive artificial variable (ξ_i). The balance between maximizing the boundary value and minimizing the misclassification errors can be controlled by identifying an edit parameter indicated by C, which takes positive values (Cortes, C. & Vapnik, V. 1995). The optimization problem for data that cannot be discriminated linearly by using the regulation parameter and the artificial variable is as in Equation 11;

$$\min\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{r}\xi_i\right] \tag{11}$$

Limitations related to this are expressed in $y_i\big((w, x_i) + b\big) - 1 \geq 1 - \xi_i$. In order to solve the optimization problem expressed in Equation 11, the data that cannot be separated linearly in the input space is displayed in a multidimensional space defined as property space (Kavzaoğlu, T. & Çölkesen, İ. 2010). Thus, the linear separation of data can be made and the hyperplane between classes can be determined. Nonlinear transformations can be made with the help of a kernel function, which is expressed as $K(X, x_i) = \varphi(X) \cdot \varphi(x_i)$ mathematically. As a result, the decision rule for the solution of a two-class problem that cannot be separated linearly using the kernel function can be written as in Equation 12 (Osuna, E. E., Freund, R. & Girosi, F. 1997);

$$f\left(X\right) = sign\left(\sum_{i=1}^{k}\alpha_i y_i \varphi\left(X\right)\varphi\left(x_i\right) + b\right) \tag{12}$$

3. EXPERIMENTAL RESULTS

350 different WBC images were used in experimental studies. These images were randomly selected and transformed into 5 different ratios of training and test data as shown in Figure 5. In addition, each dataset was randomly selected 100 times to obtain more realistic results. Thus, a total of 500 data was prepared and analyzed in statistical results for each dataset group as shown in Figure 6.

Dataset No.	Percent of dataset (*train, test*)	Number of samples		
		Training	Testing	Total
DS1	$ds(25\%, 75\%)$	87	263	350
DS2	$ds\ (33\%, 67\%)$	116	234	350
DS3	$ds(50\%, 50\%)$	175	175	350
DS4	$ds(67\%, 33\%)$	234	116	350
DS5	$ds(75\%, 25\%)$	263	87	350

Fig. 5. Datasets used in the study.

When the 5 different datasets given in Figure 5 are examined; For DS1, 25% (87 of them) of 350 WBC images were used for training and 75% (263 of them) were used for testing. Similarly, other datasets were prepared according to the specified rates. In Figure 6, all datasets are trained and tested at different rates for each machine learning algorithm; worst, best, mean, standard deviation, median and mod values are calculated.

	Dataset	Worst	Best	Mean	Std.Dev.	Median	Mod
Decision Tree	*DS1*	55,13%	74,90%	66,61%	3,62%	66,54%	65,78%
	DS2	55,98%	74,79%	67,68%	3,41%	67,95%	69,66%
	DS3	60,57%	78,86%	69,94%	3,47%	70,29%	73,71%
	DS4	56,90%	79,31%	71,17%	4,13%	71,55%	71,55%
	DS5	57,47%	**81,61%**	72,03%	4,91%	72,41%	72,41%
k-NN	*DS1*	63,50%	74,90%	69,13%	2,57%	68,82%	68,06%
	DS2	64,10%	74,79%	69,93%	2,12%	70,09%	70,51%
	DS3	64,00%	76,57%	70,53%	2,42%	70,57%	69,71%
	DS4	63,79%	77,59%	70,72%	3,51%	70,69%	75,00%
	DS5	58,62%	**81,61%**	70,89%	4,64%	71,26%	73,56%
Naïve Bayes	*DS1*	60,08%	77,19%	68,61%	3,87%	68,82%	68,82%
	DS2	61,54%	78,63%	70,17%	4,00%	70,09%	69,23%
	DS3	62,29%	80,57%	72,29%	3,53%	72,57%	72,00%
	DS4	63,79%	81,90%	73,24%	4,07%	72,41%	71,55%
	DS5	64,37%	**82,76%**	73,55%	4,21%	73,56%	72,41%
Random Forest	*DS1*	68,82%	81,37%	75,53%	3,16%	75,48%	74,52%
	DS2	69,66%	84,62%	77,07%	3,01%	77,35%	80,34%
	DS3	69,71%	85,71%	78,23%	3,42%	78,29%	78,29%
	DS4	68,97%	**91,38%**	78,94%	3,99%	79,31%	78,45%
	DS5	67,82%	90,80%	79,74%	4,66%	80,46%	80,46%
SVM	*DS1*	60,84%	77,19%	71,41%	2,52%	71,86%	72,24%
	DS2	61,97%	76,92%	72,73%	2,43%	72,86%	73,50%
	DS3	66,86%	82,29%	74,39%	2,99%	74,29%	76,57%
	DS4	66,38%	**84,48%**	74,78%	3,62%	75,00%	75,00%
	DS5	64,37%	82,76%	74,69%	4,05%	74,71%	78,16%
MLR	*DS1*	72,62%	90,11%	80,86%	3,24%	80,80%	83,65%
	DS2	76,92%	89,32%	83,97%	2,79%	83,97%	85,90%
	DS3	83,43%	94,29%	88,85%	2,23%	89,14%	89,71%
	DS4	83,62%	**96,55%**	90,17%	2,43%	90,52%	91,38%
	DS5	82,76%	**96,55%**	91,26%	2,73%	91,95%	91,95%

Fig. 6. Statistical measurements of classification success.

Figure 7 shows the box graphs showing the test success of the machine learning algorithms according to each dataset. In general, the evaluation is made; When the train and test ratios are in DS3, DS4 and DS5, more stable results are observed. Naturally, the high rate of learning is the biggest factor affecting this. When the standard deviation rates of the test success for each machine learning algorithm are examined, it is seen that MLR gives more stable results than other algorithms. To evaluate each algorithm separately according to the success of classification; According to the Decision Tree algorithm, the best result was seen in DS5 with 75.6% training and 25% for testing purposes with 81.61% classification success. It has been seen that there are 1 or 2 low outliers for all datasets. When the algorithm is evaluated according to different training and test rates, it can be said that it provides success in about 75%. According to the k-NN algorithm, the best result was seen in the DS5 with 75.6% training and 25% for testing purposes with 81.61% classification success. When the algorithm is evaluated generally according to different training and test rates, it can be said that it provides success rate of about 75% as in DT algorithm. According to Naive Bayes algorithm, the

best result was seen in DS5 with 75% training and 25% for test purposes with 82,76% classification success. When the algorithm is evaluated according to different training and test rates, it can be said that it provides success in about 75%.

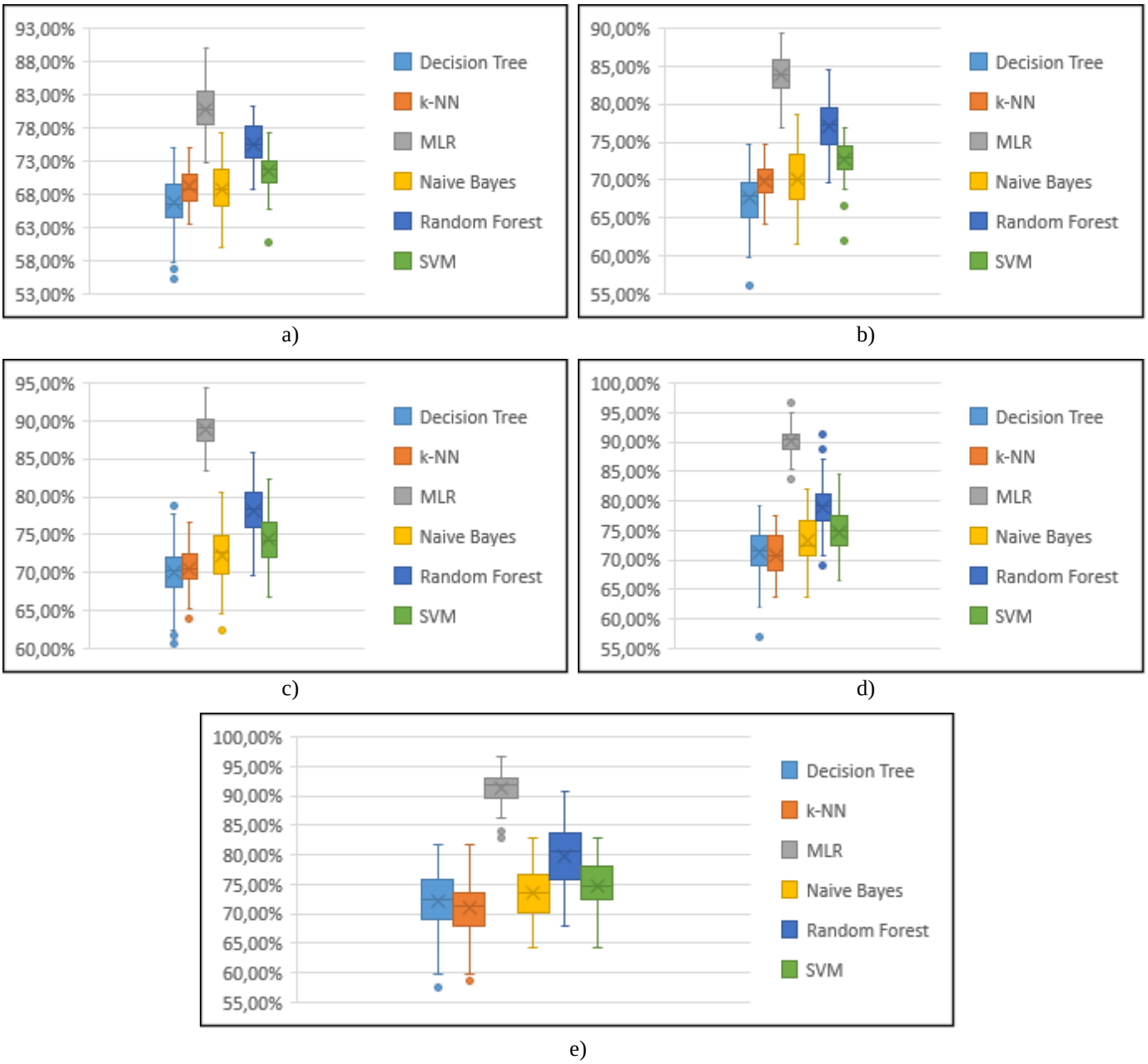

Fig. 7. Box charts of the machine learning algorithms for each dataset; a) DS1, b) DS2, c) DS3, d) DS4, e) DS5.

The best result obtained by Random Forest algorithm was seen in DS4 with 67.3% training and 33% test result with 91.38% classification success. When the algorithm is evaluated according to different train and test rates, it can be said that it provides more than 80% success. The best result obtained according to SVM algorithm was seen in DS4 with 67.4% training and 33% for testing purposes with 84.48% classification success. In cases where the training data is less than 50%, it is seen that there are 1 or 2 low outliers. When the algorithm is evaluated according to different training and test rates, it can be said that it provides success in about 80%. The best result obtained according to the MLR algorithm was observed in DS4 and DS5 datasets with 96.55% classification success. When the algorithm is evaluated according to different training and test rates, it can be said that it provides success in the range of 90% -95%. This is the greatest success rate ever achieved. Figure 8 shows the comparative result graph of all machine learning algorithms according to their best in test success. As can be seen here, the two best algorithms are MLR and Random Forest, respectively.

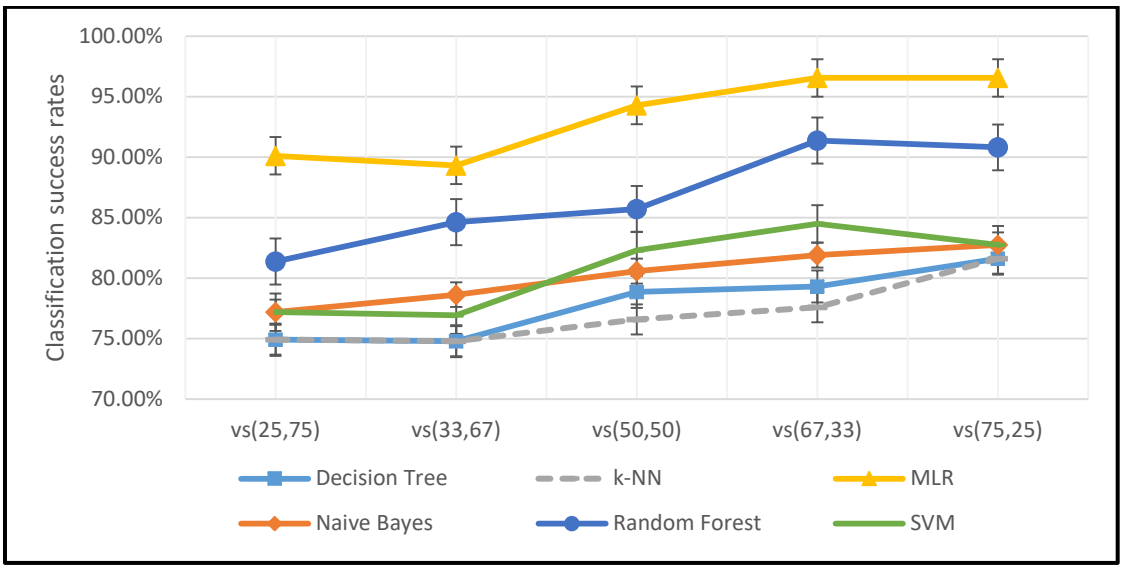

Fig. 8. The best classification performance of the machine learning algorithms.

4. CONCLUSIONS

In this study, statistical and geometrical features were extracted from microscopic blood images and a feature vector composed of 35 different parameters was formed. This feature vector is used as the input parameter for 6 different machine learning algorithms for the classification of white blood cells. In order to test the performance of the algorithms, 5 types of data were prepared in different training and test ratios, and 100 different combinations of each data-set were created and statistical results were analyzed. When the performance of classification of leukocyte cells is evaluated, it is seen that the highest success rate in all datasets and in all conditions belongs to MLR algorithm. The lowest success rate belongs to the k-NN algorithm and produced results close to the SVM and Naive Bayes algorithms. Apart from these, Random Forest algorithm is the most successful method after MLR. This method is more successful than the Decision Tree algorithm because it is a combination of more than one decision tree. As a result, the success rate of 95% obtained by the MLR algorithm is quite high and at the same time it is more stable than other methods. Therefore, the method can be applied easily for automatic classification systems. In order to further improve the classification success, the algorithm can be made more powerful by methods such as Bagging, Boosting or Bootstrapping. Thus, it is thought that global success rates can be brought to better values by reducing the factors affecting the success of blood smear images negatively.

ACKNOWLEDGEMENTS

This work was supported by research fund of the Karabük University, Project Number: KBÜ-BAP15/2-DR-003.

REFERENCES

Adjouadi, M., Zong, N. & Ayala, M. (2005). Multidimensional Pattern Recognition and Classification of White Blood Cells Using Support Vector Machines. Particle & Particle Systems Characterization, 22(2): pp. 107-118.

Arı, E., & Yıldız, Z. (2013). Parallel Lines Assumption in Ordinal Logistic Regression and Analysis Approaches. International Interdisciplinary Journal of Scientific Research, 1(3): pp. 8-23.

Avuçlu, E., & Başçiftçi, F. (2018). New approaches to determine age and gender in image processing techniques using multilayer perceptron neural network. Applied Soft Computing, 70, pp. 157–168. doi:10.1016/j.asoc.2018.05.033

Bikhet, S. F., Darwish, A. M., Tolba, H. A. & Shaheen, S. I. (2000). Segmentation and classification of white blood cells. IEEE International Conference on Acoustics, Speech, and Signal Processing, İstanbul, pp. 2259-2261.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1): pp. 5-32.

Büyüköztürk, Ş., Çokluk Bökeoğlu, Ö. & Şekercioğlu, G. (2010). Sosyal Bilimler İçin Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları. Pegem Akademi Yayıncılık, Ankara, pp. 59-65.

Cortes, C. & Vapnik, V. (1995). Support-Vector Network. Machine Learning, 20(3): pp. 273–297.

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1): 21-27.

Elen, A. & Turan, M. K. (2018). A New Approach for Fully Automated Segmentation of Peripheral Blood Smears. International Journal of Advanced and Applied Sciences, 5(1): pp. 81-93.

Ferri, M., Lombardini, S. & Pallotti, C. (1994). Leukocyte Classification by Size Functions. IEEE Workshop on Applications of Computer Vision, Sarasota, pp. 223-229.

Habibzadeh, M., Krzyzak, A. & Fevens, T. (2013). Comparative study of shape, intensity and texture features and support vector machine for white blood cell classification. Journal of Theoretical and Applied Computer Science, 7 (1): pp. 20-35.

Hiremath, P. S., Bannigidad, P. & Geeta, S. (2010). Automated Identification and Classification of White Blood Cells (Leukocytes) in Digital Microscopic Images. International Journal of Computer Applications, 2 (8): pp. 59-63.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). Applied Logistic Regression 3rd Ed. Wiley&Sons, Canada, pp. 8-35.

Jiang, Z. G., Fu, H. G. & Li, L. J. (2005). Support Vector Machine for Mechanical Faults Classification. Journal of Zhejiang University Science, 6 (5): pp. 433-439.

Joshi, M. D., Karode, A. H. & Suralkar, S. R. (2013). White Blood Cells Segmentation and Classification to Detect Acute Leukemia. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2(3): pp. 147-151.

Kavzaoğlu, T. & Çölkesen, İ. (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. Harita Dergisi, 2010(144): pp. 73-82.

Ko, B. C., Gim, J. W. & Nam, J. Y. (2011). Cell image classification based on ensemble features and random forest. Electronics Letters, 47 (11): pp. 638-639.

Krishna, P. R. & De, S. K. (2005). Naive-Bayes Classification using Fuzzy Approach. Third International Conference on Intelligent Sensing and Information Processing, Bangalore/India, pp. 61-64.

Krishnan, A. & Sreekumar, K. (2014). A Survey on Image Segmentation and Feature Extraction Methods for Acute Myelogenous Leukemia Detection in Blood Microscopic Images. International Journal of Computer Science and Information Technologies, 5 (6): pp. 7877-7879.

Krzyzak, A., Fevens, T., Habibzadeh, M. & Jelen, Ł. (2011). Application of Pattern Recognition Techniques for the Analysis of Histopathological Images. Advances in Intelligent and Soft Computing, Berlin/Germany, pp. 623-644.

Leech, N. L., Barrett, K. C. & Morgan, G. A. (2004). SPSS For Intermediate Statistics: Use and Interpretation 2nd Ed. Lawrance Erlbaum Associates Publishers, New Jersey, pp. 109-110.

Li, Q., Wang, Y., Liu, H., He, X., Xu, D., Wang, J. & Guo, F. (2014). Leukocyte cells identification and quantitative morphometry based on molecular hyperspectral imaging technology. Computerized Medical Imaging and Graphics, 38 (3): pp. 171-178.

Maji, P., Mandal, A., Ganguly, M. & Saha, S. (2015). An Automated Method for Counting and Characterizing Red Blood Cells Using Mathematical Morphology. IEEE International Conference on Advances in Pattern Recognition, Kolkata, pp. 1-6.

Orhan, U. & Adem, K. (2012). The Effects of Probability Factors in Naive Bayes Method. Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu, Bursa, pp. 722-724.

Osowski, S., Siroic, R., Markiewicz, T. & Siwek, K. (2009). Application of support vector machine and genetic algorithm for improved blood cell recognition. IEEE Transactions on Instrumentation and Measurement, 58(7): pp. 2159–2168.

Osuna, E. E., Freund, R. & Girosi, F. (1997). Support Vector Machines: Training and Applications. Massachusetts Institute of Technology and Artificial Intelligence Laboratory Report, pp. 8-10.

Pal, M. (2005). Random Forest Classifier for Remote Sensing Classification. Int. Journal of Remote Sensing, 26(1): pp. 217–222.

Pandit, A., Kolhar, S. & Patil, P. (2015). Survey on Automatic RBC Detection and Counting. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 4 (1): pp. 128-131.

Ramesh, N., Dangott, B., Salama, M. E. & Tasdizen, T. (2012). Isolation and two-step classification of normal white blood cells in peripheral blood smears. Journal of Pathology Informatics, 3 (13): pp. 1-10.

Ramoser, H., Laurain, V., Bischof, H. & Ecker, R. (2005). Leukocyte segmentation and classification in blood-smear images. IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, pp. 3371-3374.

Rawat, J., Bhadauria, H. S., Singh, A. & Virmani, J. (2015). Review of leukocyte classification techniques for microscopic blood images. 2nd International Conference on Computing for Sustainable Global Development, New Delhi, pp. 1948-1954.

Rodrigues, P., Ferreira, M. & Monteiro, J. (2008). Segmentation and Classification of Leukocytes Using Neural Networks: A Generalization Direction. Studies in Computational Intelligence, 83: pp. 373-396.

Rosin, P. L. (2003). Measuring shape: ellipticity, rectangularity, and triangularity. Machine Vision and App., 14(3): pp. 172-184.

Sanei, S. & Lee, T. K. M. (2003). Cell Recognition Based on PCA and Bayesian Classification. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, pp. 239-243.

Saraswat, M. & Arya, K. V. (2014). Automated microscopic image analysis for leukocytes identification: A survey. Micron, 65 (2014): pp. 20-33.

Saraswat, M. & Arya, K. V. (2014). Feature selection and classification of leukocytes using random forest. Medical & Biological Engineering & Computing, 52(12): pp. 1041-1052.

Sarle, W. S. (1994). Neural Networks and Statistical Models. Proceedings of the Nineteenth Annual SAS Users Group International Conference, Texas, pp. 1-13.

Sarrafzadeh, O., Rabbani, H., Talebi, A. & Yousefi-Banaem, H. (2014). Selection of the best features for leukocytes classification in blood smear microscopic images. Medical Imaging 2014: Digital Pathology, California, pp. 1-8.

Soman, K. P., Loganathan, R. & Ajay, V. (2009). Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., Delhi/India, pp. 1-10.

Sonar, S. C. & Bhagat, K. S. (2015). An Efficient Technique for White Blood Cells Nuclei Automatic Segmentation. International Journal of Scientific & Engineering Research, 6 (5): pp. 172-178.

Su, M., Cheng, C. & Wang, P. (2014). A Neural-Network-Based Approach to White Blood Cell Classification. The Scientific World Journal, 2014: pp. 1-9.

Tantikitti, S., Tumswadi, S. & Premchaiswadi, W. (2015). Image processing for detection of dengue virus based on WBC classification and decision tree. 13th International Conference on ICT and Knowledge Engineering, Bangkok, pp. 84-89.

Theera-Umpon, N. & Dhompongsa, S. (2007). Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification. IEEE Transactions on Information Technology in Biomedicine, 11(3): pp. 353-359.

Tomari, R., Wan Zakaria, Jamil, M.M.A., Nor, F.M., Fahran, N. & Fuad, N. (2014). Computer Aided System for Red Blood Cell Classification in Blood Smear Image. Procedia Computer Science, 42: pp. 206-213.

Tuceryan, M. & Jain, A. K. (1998). In the Handbook of Pattern Recognition and Computer Vision 2nd Ed. Chen, C. H., Pau, L. F. and Wang, P. S. P., World Scientific Publishing Co., pp. 207-248.

Venkatalakshmi, B. & Thilagavathi, K. (2013). Automatic Red Blood Cell Counting Using Hough Transform. IEEE Conference on Information & Communication Technologies (ICT), Thuckalay, pp. 267-271.

Washington, S. P., Karlaftis, M. G. & Mannering, F. (2003). Statistical and Econometric Methods for Transportation Data Analysis 2nd Ed. Chapman and Hall/CRC, Boca Raton/FL, pp. 263-265.

Xiong, W., Ong, S., Lim, J., Foong, K. W., Liu, J., Racoceanu, D., Chong, A. G. & Tan, K. S. (2010). Automatic Area Classification in Peripheral Blood Smears. IEEE Transactions on Biomedical Engineering, 57 (8): pp. 1982-1990.