

A LARGE-SCALE AUTOMATED CLASSIFICATION OF WHITE BLOOD CELL IMAGES

¹ALI HUSSEIN, ²FARAH SAMI, ³SHAHAB WAHHAB, ⁴DINA YOUSIF

¹Erbil Technical Engineering College / Erbil Polytechnic University, Kirkuk Road, Erbil, Iraq

²Erbil Technical Engineering College / Erbil Polytechnic University, Kirkuk Road, Erbil, Iraq

³Erbil Technical Engineering College / Erbil Polytechnic University, Kirkuk Road, Erbil, Iraq

³College of Engineering and Computer Science/ Lebanese French University, Erbil, Iraq

⁴Erbil Technical Engineering College / Erbil Polytechnic University, Kirkuk Road, Erbil, Iraq

E-mail: ¹ali.yousif@epu.edu.iq, ²farah.xoshihi@epu.edu.iq, ³shahab.kareem@epu.edu.iq, ⁴dina.mikhail@epu.edu.iq

Abstract - The counts of several types of white blood cells provide valuable information for diagnosing many diseases. The automation of this task saves time and avoids errors in counting. In this paper, we attempt to classify the white blood cells in peripheral blood based on the shapes and the features of the nucleus. We implement a system and use it to identify and analyze the White Blood Cells (WBCs) automatically. The proposed system can be applied in four steps, namely, segmentation, scanning, feature extraction, and classification of a blood cell. First, we segment the cell images, which involve the categorization of white blood cells into clusters. The second step entails the scanning of each segmented image and preparing the dataset. Extracting the shape and texture from a scanned image is the third step. In the final stage, we apply different machine learning algorithms (SVM, Random Tree, Zero-R) to classify the result based on these criteria.

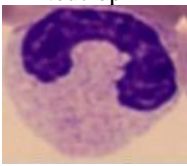
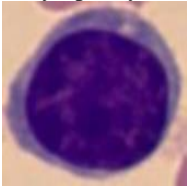
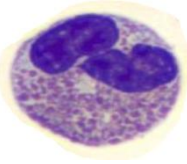
Keywords - Machine learning (ML), Segmentation, Digital image, Image extraction, Histogram.

I. INTRODUCTION

White blood cells - or leukocytes - protect the body against infectious disease. These cells are colorless, but we can use special stains on the blood that make them colored and visible under the microscope.

White blood cells are the largest blood cell and can move by sticking out one part of their body and dragging the rest of themselves along. They are the "soldiers" of the blood, attacking bacteria and other invaders unfamiliar to the body. White blood cells can squeeze through tiny blood vessels, leaving the bloodstream to enter other tissues that are being

attacked by foreign invaders. Most white blood cells are manufactured in the red marrow of bones. Some are also made in special glands elsewhere in the body. In a healthy person, there are between 4 and 11 thousand leukocytes in every cubic inch of blood [1], [2], [3]. When a person has an infection, this signals the marrow and special glands to make more white blood cells [4]. When a medical technologist counts the white blood cells in someone's blood, they can tell the doctor if there is an infection. There are five types of white blood cells. They are neutrophils, eosinophils, basophils, lymphocytes, and monocytes Table 1.

Type of white blood cell	Function and Description
<p>Neutrophil</p> 	<p>It helps stop microorganisms in infections by eating them and destroying them with enzymes.</p>
<p>Lymphocyte</p> 	<p>While they are smaller compared to other leukocytes, lymphocytes have a large round nucleus that takes up much of the cell volume. As a result, the lymphocytes have extraordinarily little to no cytoplasm and uses antibodies to stop bacteria or viruses from entering the body (B-cell lymphocyte) also kills off the body's cells if they have been compromised by a virus or cancer cells (T-cell lymphocyte).</p>
<p>Eosinophils</p> 	<p>Compared to neutrophils that may have 2 to 5 lobed nuclei, eosinophils only have a bi-lobed (two lobes) nucleus that is shaped like a horse-shoe. They will also appear spherical with fine granules referred to as acidophilus refractive granules. It helps control inflammation, especially active during parasite infections and allergic reactions, stops substances or other foreign materials from harming the body.</p>

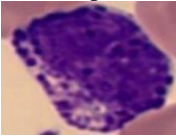
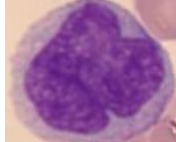
<p style="text-align: center;">Basophils</p> 	<p>Compared to the other granulocytes, basophils have a large and irregular nucleus that resides inside the spherical shaped cell. Whereas the nucleus of the other granulocytes is well defined and can be clearly described, the nucleus of a basophil (bluish under the microscope) is large and irregular inside the cell and may prove difficult for researcher to describe. Basophils are responsible for produces enzymes during asthma attacks and allergic reactions.</p>
<p style="text-align: center;">Monocytes</p> 	<p>Compared to lymphocytes (a granular leukocyte) monocytes are larger with a nucleus that is a bean or kidney-shaped. These cells also have more cytoplasm compared to lymphocytes it becomes a macrophage in the body's tissues, eating microorganisms and getting rid of dead cells while increasing immune system strength.</p>

Table.1 Types of WBCs and their functions

II. IMAGE PREPROCESSING

Through the expanding application of direct digital imaging systems to medical diagnostics, digital image processing becomes more and more necessary in health care [5]. Real-world data is usually incomplete and noisy and also is expected to include unnecessary and repetitive information or errors [6]. Also, remarkable standard methods are extremely susceptible to these foretellers, like linear regression. Therefore, analysis including preprocessing data before beginning the model is required. This section outlines some important methods in data preprocessing, including data cleaning, data transformation, and data modulation. Data preprocessing in common than usually consist of the conversion from the collection of attributes to another collection of attributes to allow the appropriate data mining or machine learning technique to obtain better results [7]. Classifier development is one of the commonly researched issues in data mining and machine learning areas. Thousands of algorithms have been submitted. The quality of the learned models, but, based on the essence of the training data. Neither thing which classifier inducer is applied, if the training data are unreliable, bad models will result [8]. Image visualization applies to all kinds of manipulation of this matrix, producing an optimized production of the image. Image analysis holds all the levels of processing, which is applied for quantitative measurements as well as abstract interpretations of biomedical images.

These actions require a priori information on the nature and content of the images, which must be integrated into the algorithms on a high level of abstraction. Thus, the method of image analysis is extremely specific, and enhanced algorithms can be carried unusually quickly into other application areas. Image administration sums up all methods that present efficient warehouse, communication, archiving, transmission, and access (retrieval) of image data. Thus, the methods of telemedicine are also a section of the image administration [9]. Indifference to image analysis, which is usually further pointed to as high-level image processing,

low-level processing means standard or automatic procedures, which can be achieved without a priori information on the exclusive content of images. Figure 1(a) shows a representative microscopic image of a human blood cell with four WBCs also several RBCs [10]. In blood cell image exposure, the task is normally existing image enhancement, for the objective of reducing noise. Cell segmentation requires the elimination of the background containing red blood cells, platelets, and different things from the image obtained. White blood cells, the objects of power, grow as the product of the segmentation process. Accurate segmentation should yield the complete white blood cell, including the nucleus and the cytoplasm. The shape of the nucleus, its texture, area, and the ratio of its content in the cell are some of the features that are needed to classify the cell. Image segmentation is the most significant step and a principal technology in image processing, and it will instantly change succeeding processing [11], [12]. To scientific theories, image segmentation has performed excellent development and a lot of novel segmenting algorithms have been introduced. However, the largest algorithms have their drawbacks. As for cell images, owing to the complicated universe, it nevertheless contains a challenging responsibility to segment and count them [13]. There are various statements of Morphological Processing (MP) in different areas of biomedical image processing. Noise is decreased, smoothing, and other types of filtering, classification, segmentation, and pattern recognition are utilized to both binary and grayscale images.

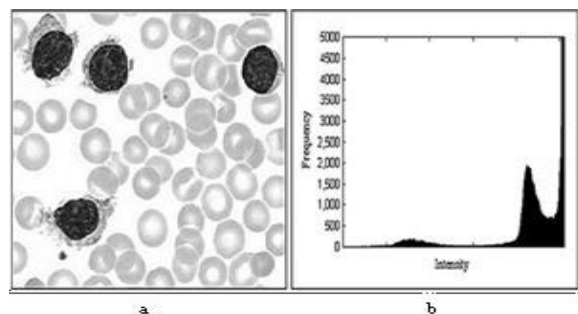


Figure 1: (a) Microscopic image of a human blood cell
(b) Histogram of the image

III. CLASSIFICATION ALGORITHMS

Machine learning (ML) is an algorithm set particularly agreed to forecast. Those ML techniques are simpler to perform and also offer better than the standard mathematical approaches [14], [15]. Classification is a machine learning problem on how to select numbers on distinct data depending on a given set of specified data. The classification techniques include predicting a confident result depending on a given input. To predict the result, the technique processes a training collection, including a collection of properties and a particular result, normally announced object or forecast characteristic. The classification technique selects pixels in the image to sections or categories of concern. There are a couple of models of classification algorithms supervised, and unsupervised. Supervised classification utilizes the phantom marks received from training samples unless data to analyze a dataset or image. Unsupervised classification identifies spectral classes in a multiband image without the analyst's invasion. The Image Classification algorithms support unsupervised classification through producing technology to build the clusters, the capability to investigate the quality of the clusters, and passage to classification algorithms. There are various algorithms enhanced by researchers across the years. To classify a set of data into various groups or classes, the correlation between the classes and the data within which they are classified is necessary to be completely known. In this paper, three algorithms will be presented. Classification of cells is more important in the medical image [16]. The mathematical model behind these algorithms is illustrated in the next section.

IV. ALGORITHMS AND EXPERIMENTS

1. Methodology:

The step of pre-processing regularly involves improving the image from the image collected. Furthermore, this step implies implementation within the sequence for producing the primary image further fitting to the next estimate. After preprocessing we start the segmentation step. In these parts, we apply different steps starting from reading image color blood, then converting it to the grayscale image. The automatic detection and classification of white blood cells is an innovative technology. Matlab2016a has been used to simulate the prototype. In this paper, the proposed system involves the following three stages : (1) Segmentation and Scanning WBCs from a blood smear image, (2) Feature extraction to obtain the database, and (3) Classification of blood cells into one of five classes (Figure 2).

In the following paragraphs, we briefly explain each of these steps.

The performance of an automatic white blood cell classification system depends on a good segmentation

algorithm for segmenting white blood cells from other blood smear components. A good segmentation can be obtained where:

- Pixels that are in the same category share similar multivariate grayscale values and form a connected region
- Neighboring pixels that belong to various categories have various values. Segmentation is considered as the essential step in analyzing WBC image: where working becomes on objects (containing many pixels) in the image instead of observing single pixels.

Feature extraction has morphological operations, where it does the role of extracting features out of WBCs that contain vital information. The shape feature includes the area of the nucleus and the whole cell. Texture features include homogeneity, contrast, and entropy. They are extracted from resulted segmentation images.

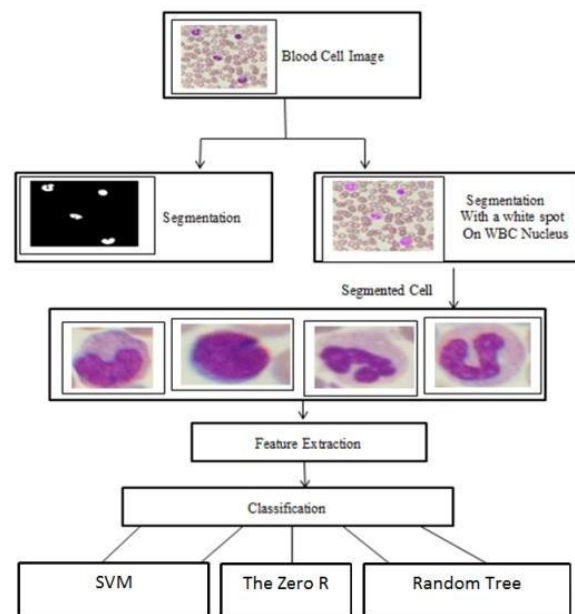


Figure 2: Block Diagram of Proposed Method

2. Results and Discussion:

Before presenting and discussing the result, we start to present the criteria (Correlation coefficient, mean absolute error, Root-mean-square deviation, Relative absolute error, and Root relative squared error) that used to present the results. The first criterion is the correlation coefficient.

A correlation coefficient measures the strength and direction of a linear association between two variables. It ranges from -1 to 1. The closer the absolute value is to 1, the stronger the relationship. A correlation of zero indicates that there is no linear relationship between the variables [17].

Mean absolute error (MAE) is an error statistic that averages the distances between each pair of actual (Z_t) and fitted forecast (Z_t^*) data points. MAE is calculated by taking the average of the absolute

errors. In addition, it is most appropriate when the cost of forecast errors is proportional to the absolute size of the forecast errors. MAE is given by:

$$MAE = \frac{1}{N} \sum_t^N |Z'_t - Z_t| = \frac{1}{N} \sum_t^N |e_t| \quad (1)$$

Suppose, (e_1, t, e_2, t) , $t = 1, 2, \dots, m$ are hi-step, out-of-sample forecast errors of models 1 and 2, respectively. Taking MAE as a measure of prediction loss, the loss differential from the two models can be expressed as date

$$= |e_1 t| - |e_2 t|, \text{ to } = 1, 2, \dots, m \text{ [18],[19].}$$

The root means square deviation (RMSD) (also called the Root Mean Square Error (RMSE)) is a frequently used measure of the difference between values predicted by a model and the values observed from the environment that is being modeled. These individual differences are also called residuals, and the RMSD serves to aggregate them into a single measure of predictive power. The RMSD of a model prediction for the estimated variable X model is defined as the square root of the mean squared error:

$$RMSD = \sqrt{\frac{\sum_{i=1}^n (X_{obsy,i} - X_{model,i})^2}{n}} \quad (2)$$

Jobsere X_{obs} has observed values and X model is modeled values at a time/place I. The calculated RMSE values will have units, and the RMSD for phosphorus concentrations can, for this reason, not be directly compared to RMSE values for chlorophyll-a concentration, etc. However, the RMSD values can be used to distinguish model performance in a calibration period from that of a validation period as well as to compare the individual model performance to that of other predictive models. RMSD is the square root of the average of squared errors. The effect of each error on RMSD is proportional to the size of the squared error; thus, larger errors have a disproportionately large effect on RMSD. Consequently, RMSD is sensitive to outliers [20], [21].

Relative absolute error (RSE) is the different degree between the absolute deviation obtained from the prediction model and the absolute deviation obtained by directly speculating the training sample. It is inversely proportional to prediction accuracy. The smaller the RSE is, the higher the prediction accuracy can be:

$$RSE = \frac{\sum_{i=1}^n |f_i - y_i|}{\sum_{i=1}^n |f'_i - y_i|} \quad)3($$

RRSE (root relative squared error) can be calculated as follows:

$$RRSE = \frac{\sum_{i=1}^n |f_i - y_i|^2}{\sum_{i=1}^n |f'_i - y_i|^2} \quad)4($$

RRSE is also inversely proportional to prediction accuracy. The smaller the RRSE is, the higher the prediction accuracy can be [22]. The microscopic images used in this paper are five types of WBC

images. The image for WBC was taken from the central public health Laboratory in Duhok. Images are captured from smear slides by a Nikon 50i microscope, equipped with a Nikon color camera DP5M.

After applying the mentioned classification algorithm in section III, we present the result of different criteria such as shown the result Figures 3-7.

The output of the classification algorithms is classified into five different models (Basophil, Eosinophil, Lymphocyte, Monocyte, and Neutrophil). The result of the classification is shown in the Figures below.

Depending on the Correlation coefficient shown in figure 3, Random Tree is the best algorithm. While depending on the Mean Absolute Error shown in figure 4, Root means square error in figure 5, Relative absolute error in figure 6, and Root relative squared error in figure 7, the Zero R classifier is better than the other mentioned algorithms.

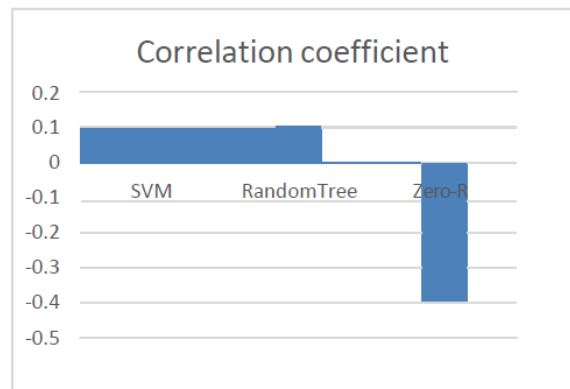


Figure 3: Correlation coefficient of mentioned algorithms

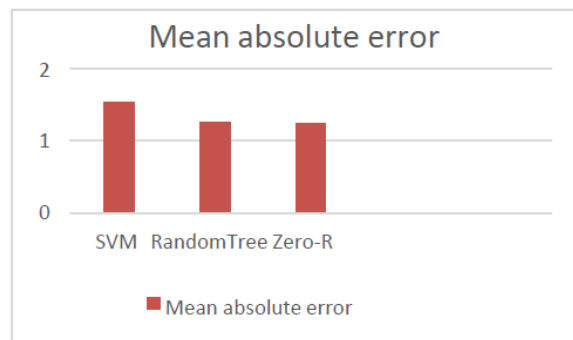


Figure 4: Mean Absolute error of mentioned algorithms

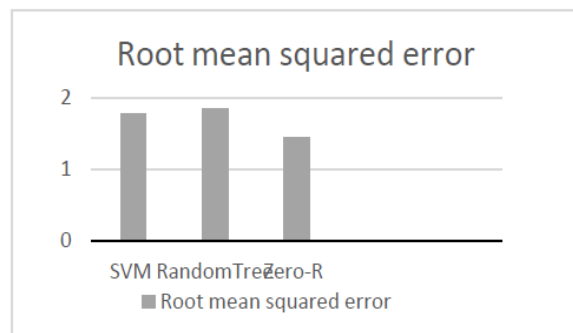


Figure 5: Root mean square error of mentioned algorithms

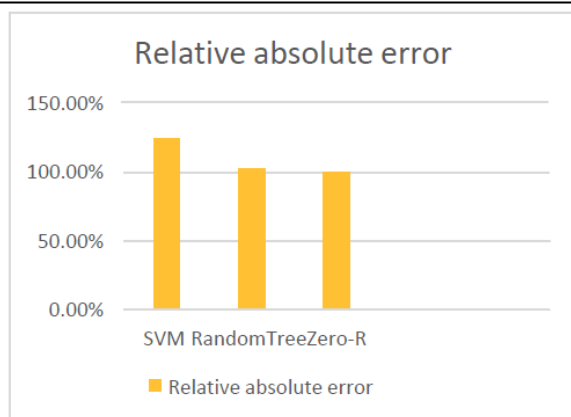


Figure 6: The relative absolute error of mentioned algorithms

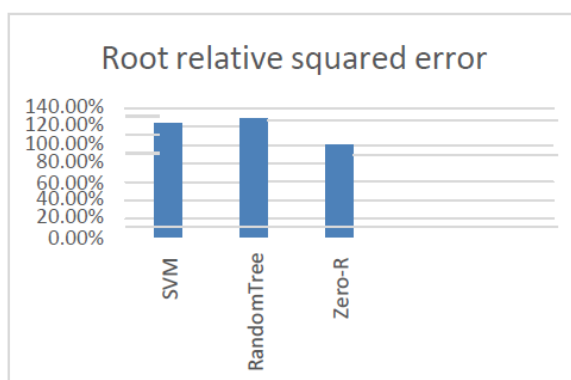


Figure 7: Root relative squared error of mentioned algorithms

V. CONCLUSION

During this article, we proposed a process for specific identification of WBCs also bits depending on image classification, improvement, and segmentation. Different existed methods have been utilized to perform specific classification steps. We used (SVM, Random Tree and the Zero R classifier) classification methods for classifying the WBC images into five parts (Basophil, Eosinophil, Lymphocyte, Monocyte, and Neutrophil), then used five types of evaluation metrics to find the best one. The result showed that, the Zero R classifier is the best of the mentioned algorithms in this dataset, depending on the Mean Absolute Error, Root means square error, Relative absolute error, and Root relative squared error evaluation metrics, while Random Tree classifier is the best depending on the Correlation coefficient evaluation metrics.

REFERENCES

- [1] MINAL, D. J., Atul H. K. And SURALKAR, R. S. (2013) White Blood Cells Segmentation and Classification to Detect Acute Leukemia. Vol. 2, Issue 3. Maharashtra: International Journal of Emerging Trends & Technology in Computer Science.
- [2] Jaroornrut Prinyakupt and Charnchai Pluempitiwiriyaewej, "Segmentation of white blood cells and comparison of cell

morphology by linear and naïve Bayes classifiers", BioMed Eng OnLine (2015).

- [3] Anjali Gautam; Harvindra Bhadauria, "Classification of white blood cells based on morphological features", International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014.
- [4] Sadat Nazlibilek, Deniz Karacor, Korhan Levent Ertürk, Gokhan Senegal, Tuncay Ercan, Fuad Aliew, "White Blood Cells Classifications by SURF Image Matching, PCA and Dendrogram",
- [5] Siddhartha Banerjee, Bibek Ranjan Ghosh, Surajit Giri, and Dipayan Ghosh, "Automated System for Detection of White Blood Cells in Human Blood Sample", Springer Nature Singapore Pte Ltd. 2018.
- [6] Samir K. Bandyopadhyay, "METHOD FOR BLOOD CELL SEGMENTATION", Journal of Global Research in Computer Science, Volume 2, No. 4, April 2011.
- [7] Kuhn, Max, and Kjell Johnson. Applied predictive modeling. New York: Springer, 2013.
- [8] Quilumba, Franklin L., Et al. "An overview of AMI data preprocessing to enhance the performance of load forecasting." Industry Applications Society Annual Meeting, 2014 IEEE. IEEE, 2014.
- [9] Quinlan J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
- [10] Tan P.-N., Steinbach M., Kumar V.: Introduction to Data Mining. Addison Wesley, Boston, MA, 2006.
- [11] Faisal Kamiran • Toon Calders, "Data preprocessing techniques for classification without discrimination", Knowl Inf Syst. (2012) 33:1–33 DOI 10.1007/s10115-011-0463-8.
- [12] Thomas M. Deserno, "Biomedical Image Processing", Springer- Verlag Berlin Heidelberg 2011.
- [13] HUIYU Z., JIAHUA, W. and JIANGUO, Z. "Digital Image Processing Part II.", backbone, London, 2014.
- [14] Sehla Loussaief, Afef Abdelkrim, "Machine Learning Framework for Image Classification", 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2016.
- [15] Tanmoy Das, "Machine Learning algorithms for Image Classification of hand digits and face recognition dataset", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 04 Issue: 12 | Dec-2017.
- [16] Deeman Y. Mahmood, Dr. Mohammed A. Hussein, "Intrusion Detection System Based on K-Star Classifier and Feature Set Reduction", IOSR Journal of Computer Engineering (IOSR- JCE) Volume 15, Issue 5 (Nov. - Dec. 2013).
- [17] Breiman, L.: Bagging predictors. Technical Report 421, Department of Statistics, University of California at Berkeley, 1994.
- [18] S. W. Kareem, "An Evaluation ALgorithms for Classifying Leukocytes Images," 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic" (IEC), 2021.
- [19] Kristína Machová, František Barčák, Peter Bednár, "A Bagging Method using Decision Trees in the Role of Base Classifiers", Acta Polytechnica Hungarica Vol. 3, No. 2, 2006.
- [20] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse, "WEKA Manual for Version 3-7-8", University of Waikato, Hamilton, New Zealand, 2013.
- [21] Dr. Roggenbach and Prof. Schlingloff, "A Review Paper on Decision Table-Based Testing", Cai Ferriday 345399, January 7th, 2007.
- [22] G. Asuero, A. Sayago, and A. G. Gonz´alez, "The Correlation Coefficient: An Overview", Critical Reviews in Analytical Chemistry, Taylor and Francis Group, LLC,36:41– 59, 2006.

