Computer-aided classification of images containing white blood cells

Ali Hussein Yousif[a,*], Shahab Wahhab Kareem[b,a], Dina Yousif Mikhail[a], Farah Sami Khoshaba[a]

[a]*Department of Information System Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Kirkuk Road, Erbil, Iraq*
[b]*Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, Erbil, Iraq*

(Communicated by Asadollah Aghajani)

Abstract

The counts of various types of white blood cells offer important information that can be used in the diagnostic process for a wide variety of disorders. The automation of this procedure allows for time savings and eliminates the possibility of counting mistakes. In this study, the authors make an attempt to categorize the white blood cells that are found in the peripheral blood based on the shapes of the nuclei and the characteristics that they exhibit. The authors put in place a system and make use of it to automatically identify and analyze White Blood Cells (WBCs). A blood cell can be segmented, scanned, have its features extracted, and then be classified using the system that was proposed. These are the four processes that make up the process. To begin, the authors used segmenting the cell images, which involves grouping white blood cells into their respective clusters. The second part of the process consists in scanning each image that has been segmented and producing the dataset. The third phase involves the form and texture of an image that has been scanned. In the final stage, the authors apply various machine-learning techniques to classify the outcome based on these criteria. These methods include Naïve Bayes, Random Tree, and K-star.

Keywords: Machine learning (ML), Segmentation, Digital image, Image extraction, Histogram
2020 MSC: 62H35, 65D18, 94A08, 97P80

1 Introduction

White blood cells, also known as leukocytes, are responsible for defending the body against bacterial and viral infections. These cells do not have any color of their own, but by applying specific stains to the blood, we can give them color and make them visible under a microscope.
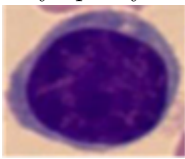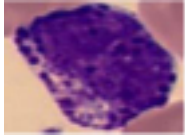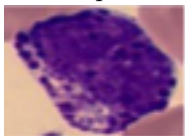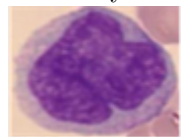
The white blood cell is the largest type of blood cell and has the ability to move by protruding one part of its body and dragging the rest of itself along behind it. They are the "soldiers" of the blood, and their job is to fight off germs and other foreign invaders that the body does not recognize. White blood cells can pass through narrow blood channels, which allows them to exit the circulatory system and travel to other tissues that are under attack from outside pathogens. The majority of white blood cells are produced in the bone marrow, often known as the red marrow. Some are also produced in specialized glands located in different parts of the body. There are between 4,000

---

*Corresponding author
Email addresses: `ali.yousif@epu.edu.iq` (Ali Hussein Yousif), `shahab.kareem@epu.edu.iq` (Shahab Wahhab Kareem), `dina.mikhail@epu.edu.iq` (Dina Yousif Mikhail), `farah.xoshibi@epu.edu.iq` (Farah Sami Khoshaba)

and 11,000 leukocytes packed into every cubic inch of blood in a healthy person [11, 21, 9]. When a person gets an infection, the bone marrow and other specific glands in the body receive a signal to produce more white blood cells [20]. If there is an infection in a patient, a medical technologist who counts the white blood cells in that person's blood can let the attending physician know about it. There are five different kinds of white blood cells in the human body. These white blood cells, known as neutrophils, eosinophil, basophils, lymphocytes, and monocytes, are shown in Table 1.

Table 1: Types of WBCs and their functions

Type of white blood cell	Function and Description
Neutrophil	It does this by digesting the germs that cause infections and then using enzymes to kill them completely.
Lymphocyte	Lymphocytes, in comparison to other types of leukocytes, are smaller; nonetheless, their nuclei are significantly larger and rounder, and they occupy the majority of the cell's volume. As a consequence of this, lymphocytes have extremely little to no cytoplasm, and they use antibodies to prevent bacteria or viruses from entering the body (in the case of B-cell lymphocytes). Lymphocytes also eliminate cells in the body that have been infected by a virus or cancer cells if these cells have become compromised (T-cell lymphocyte).
Eosinophils	Eosinophils, on the other hand, only have a bi-lobed (two lobes) nucleus that is shaped like a horseshoe, in contrast to neutrophils, which can have anywhere from 2 to 5 lobed nuclei. Additionally, they will have a spherical appearance and will have fine granules that are referred to as acidophilus refractive granules. It is highly active during parasite infections and allergic reactions, and it helps manage inflammation. Additionally, it prevents harmful substances and other foreign elements from causing harm to the body.
Basophils	Basophils are distinguished from other types of granulocytes by their big and atypically shaped nuclei, which are located within their spherical-shaped cells. In contrast to the nuclei of the other granulocytes, which are well defined and can be described in great detail, the nucleus of a basophil (which appears bluish under the microscope) is vast and irregular inside the cell, which may make it challenging for researchers to explain. During episodes of asthma and allergic responses, basophils are the cells that are responsible for producing enzymes.
Monocytes	Monocytes are granular leukocytes that are larger than lymphocytes and have a nucleus that is kidney- or bean-shaped. Lymphocytes are another type of leukocyte. In comparison to lymphocytes, these cells have a greater amount of cytoplasm. When they enter the tissues of the body, they transform into macrophages, which consume harmful germs, eliminate dead cells, and boost the function of the immune system.

2 Image Preprocessing

Digital image processing is becoming increasingly important in the medical field as a result of the growing use of direct digital imaging systems for diagnostic purposes [4]. Data collected from the real world are typically noisy and lacking, and it is to be expected that they also contain information that is redundant, pointless, or both [3]. In addition, remarkable traditional approaches, such as linear regression, are exceptionally vulnerable to the influence of these predictors. As a result, it is necessary to do an analysis that involves preparing the data before initiating the model. This section provides an overview of some essential procedures involved in the preprocessing of data. These procedures include data cleaning, data transformation, and data modulation. Data preparation in common than typically consist of the conversion from one group of attributes to another collection of attributes in order to enable the proper data mining or machine learning technique to produce better results [14]. The creation of classifiers

is one of the topics that receive a lot of attention from researchers in the fields of data mining and machine learning. There have been thousands of algorithmic solutions offered. The quality of the models that were learned, but with an emphasis on the core aspects of the training data. If the training data are untrustworthy, then it does not matter which classifier inducer is applied; the end result will still be inaccurate models [22]. Image visualization is applicable to any and all forms of manipulating this matrix, which results in the image being produced in the most effective manner. Image analysis encompasses all levels of processing and can be used to perform quantitative measurements as well as interpret biomedical pictures in a more general sense.

To carry out these activities, a priori knowledge regarding the structure and subject matter of the images is required. This knowledge must be incorporated into the algorithms at a level of abstraction that is sufficiently high. Because of this, the method of image analysis is particularly specialized, and improved algorithms can be transferred into other application areas in an abnormally short amount of time. The term "image administration" refers to the collection of practices that provide the storage, communication, archiving, transmission, and access (retrieval) of image data in an effective manner. Therefore, the processes involved in telemedicine are likewise considered to be a part of image administration [23]. Low-level processing, also known as standard or automatic procedures, refers to image processing that is not concerned with image analysis, which is typically referred to as high-level image processing. These procedures can be completed without requiring any a priori knowledge regarding the specific content of images. [24] The image shown in Figure 1(a) is a representative microscopic image of a human blood cell that contains numerous RBCs as well as four WBCs. The mission of blood cell image exposure is often existing image enhancement, with the goal of noise reduction as the primary focus of the work. Eliminating the backdrop from the image produced, which may contain various items such as red blood cells and platelets, is necessary for cell segmentation. The method of segmentation results in the growth of white blood cells, the most powerful things in the body. When segmentation is done correctly, the whole white blood cell should be produced, containing both the nucleus and the cytoplasm. Some of the characteristics that are required to classify a cell include the form of the nucleus, as well as its texture, area, and the ratio of its content to the total volume of the cell. Image segmentation is the most important stage and a primary technique in image processing. It will immediately impact the processing that comes after it [12, 8]. Image segmentation has made tremendous strides for the advancement of scientific theories, and a great number of innovative segmenting algorithms have been developed. However, even the most comprehensive algorithms have their limitations. When it comes to cell pictures, the job of segmenting and counting them can be difficult due to the complexity of the cosmos [25]. The term "Morphological Processing" (MP) can refer to a number of different things depending on the area of biomedical image processing you're looking at. Both binary and grayscale images are subjected to noise reduction, smoothing, and many forms of filtering, classification, and segmentation techniques. Pattern recognition is also performed on these image types.
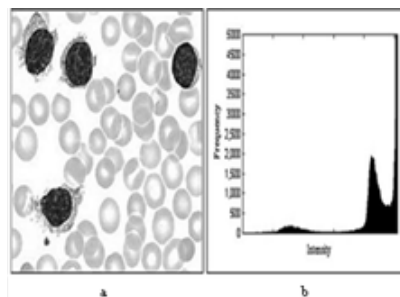

Figure 1: (a) Microscopic image of a human blood cell (b) Histogram of the image

3 Classification Algorithms

Machine learning (ML) is an algorithm set particularly agreed to forecast. Those ML techniques are simpler to perform and offer better than the standard mathematical approaches [15, 7]. 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 development, the method processes a training collection, including a collection of properties and a particular outcome, typically an 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 by producing technology to build the clusters, the capability to investigate the quality of the collections, 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 [17]. The mathematical model behind these algorithms is illustrated in the next section.

4 Algorithms And Experiments

4.1 Methodology

The stage known as "pre-processing" typically entails making improvements to the image that was obtained from the previous step. In addition, this step necessitates execution inside the sequence for the purpose of establishing the primary image and subsequently fitting it to the subsequent estimate. Following the completion of the preprocessing step, we go on to the segmentation step. In these parts, we apply different steps starting from reading the image color blood, then converting it to a grayscale image. The automatic detection and classification of white blood cells is an innovative technology. Matlab 2020a was utilized in order to carry out the simulation of the prototype. The three stages that will be discussed in this paper as part of the proposed system are as follows:

The first step is the segmentation and scanning of WBCs from a blood smear picture, followed by the second step, which is the extraction of features in order to obtain the database, and the third step is the classification of blood cells into one of five classes (Figure 2). In the paragraphs that follow, we will provide a concise explanation for each of these stages. The efficiency of an automatic white blood cell classification system is directly proportional to the quality of the segmentation algorithm used to separate white blood cells from the other components of a blood smear. When these conditions are met, a satisfactory segmentation can be obtained. Pixels that belong to the same category have the same grayscale values across all of their variables and create a linked region. Pixels that are adjacent to one another and fall into a variety of categories each have a unique value. In the process of evaluating a WBC image, the segmentation stage is considered to be the most important one. During this step, the focus shifts from monitoring individual pixels to working with individual objects inside the image. In morphological operations, feature extraction plays the job of extracting features from WBCs that hold essential information. These WBCs can be thought of as information storage containers. The region of the nucleus, as well as the complete cell, are both included in the shape feature. Homogeneity, contrast, and entropy are all characteristics of a texture's appearance. They are taken from the images that were produced as a result of the segmentation process.
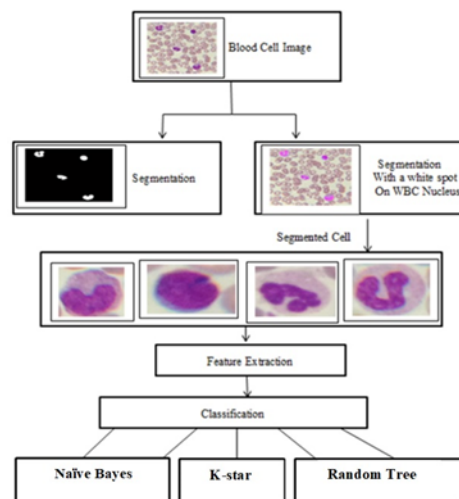

Figure 2: Block Diagram of Proposed Method

4.2 Results and Discussion

Before presenting and discussing the results, we will first offer the criteria that were utilized to show the results. These criteria include the correlation coefficient, mean absolute error, Root-mean-square deviation, Relative absolute

error, and Root relative squared error. The correlation coefficient is the primary criterion to look at. A correlation coefficient is a statistical tool that determines the degree of linear link between two variables as well as the direction of that association. It can go as low as -1 or as high as 1. The greater the proximity of the absolute value to 1, the more significant the association. If there is no linear relationship between the variables, then you will have a correlation of zero [6].

The mean absolute error, also known as MAE, is a type of error statistic that takes the average distance between each pair of actual data points (Zt) and fitted prediction data points (Z't). To determine MAE, simply take the average of all of the absolute mistakes in the calculation. In addition to this, it is most suitable for use in situations in which the cost of incorrect forecasts is proportional to the absolute magnitude of those forecasts. What determines MAE is:

$$MAE = \frac{1}{N} \sum_{t}^{N} |Z'_t - Z_t|) = \frac{1}{N} \sum_{t}^{N} |e_t| \tag{4.1}$$

Let's say that (e1, t, e2, t), where t= 1, 2,...m are the high-step, out-of-sample forecast errors of model 1 and model 2, respectively. Using the mean absolute error as a measure of the loss in accuracy of the forecast, the loss differential between the two models may be stated as date = | e1t |-| e2t |, where to = 1, 2,...m [13, 10].

The root mean square deviation (RMSD), which is also known as the root mean square error (RMSE), is a measurement that is used quite frequently to determine the degree of dissimilarity that exists between the values that are predicted by a model and the values that are observed from the environment that is being modeled. These individual differences, which are sometimes referred to as residuals, are compiled into a single measure of predictive power by using the root-mean-squared deviation (RMSD). The root square of the mean squared error is the definition of what is known as the root mean squared deviation (RMSD) of a model prediction for an estimated variable X model: 4.2

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs\rightleftharpoons,i} - X_{model,i})^2}{n}} \tag{4.2}$$

Xobs contains values that have been observed, and X model represents values that have been modeled at a certain time and location I. The RMSE values that are calculated will have units, and because of this, the RMSD values for phosphorus concentrations cannot be directly compared to the RMSE values for chlorophyll-a concentration, or any other concentration, for that matter. However, the RMSD values can be utilized to differentiate between the performance of a model during a calibration period and that of a validation period. Additionally, these values can be utilized to compare the performance of an individual model to that of other predictive models. The root square of the average of squared mistakes is the value known as RMSD. Since the influence of each error on RMSD is proportional to the size of the squared error, larger errors have a disproportionately significant effect on RMSD. RMSD is defined as the root mean square deviation. Because of this, RMSD is susceptible to the presence of outliers [5, 19, 18, 1, 2, 16].

RSE stands for "relative absolute error," which refers to the degree to which the absolute deviation produced from the prediction model differs from the absolute deviation acquired by simply speculating on the training sample. It has a negative relationship with the accuracy of the prediction. When the relative standard error (RSE) is lower, the forecast accuracy can be higher: 4.3

$$RSE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{\sum_{i=1}^{n} |f'_i - y_i|\prime} \tag{4.3}$$

The formula for calculating RRSE, also known as root relative squared error, is as follows: 4.4

$$RRSE = \frac{\sum_{i=1}^{n} |f_i - y_i|^2}{\sum_{i=1}^{n} |f'_i - y_1|\prime^2} \tag{4.4}$$

The RRSE also has an inverse relationship with the accuracy of the prediction. The RRSE should be as low as possible for the forecast accuracy to be as high as possible [18]. This article makes use of five distinct types of WBC pictures for its microscopic illustrations. The photograph that was used for WBC was obtained from the central public health Laboratory in Duhok. Smear slides are examined with a Nikon 50i microscope that is fitted with a Nikon color camera DP5M for the purpose of image acquisition.

After applying the described classification method in section III, we offer the results of various criteria, such as those displayed in Figures 3-7, which are the outcome of applying the algorithm. The results of the categorization algorithms are separated into five distinct models (Basophil, Eosinophil, Lymphocyte, Monocyte, and Neutrophil). The outcome of the classification is presented in the Figures that can be seen below.

The best algorithm is the K-star, which can be determined by looking at figure 3, which shows the correlation coefficient. The K-star classifier performs significantly better than the other algorithms that were mentioned, depending on the Mean Absolute Error that is displayed in figure 4, the Root Means Square Error that is displayed in figure 5, the Relative Absolute Error that is displayed in figure 6, and the Root Relative Squared Error that is displayed in figure 7.
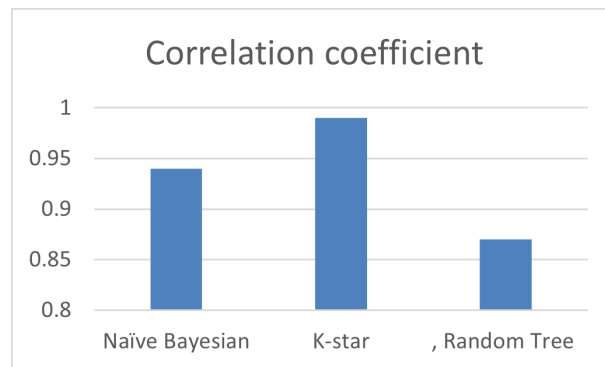

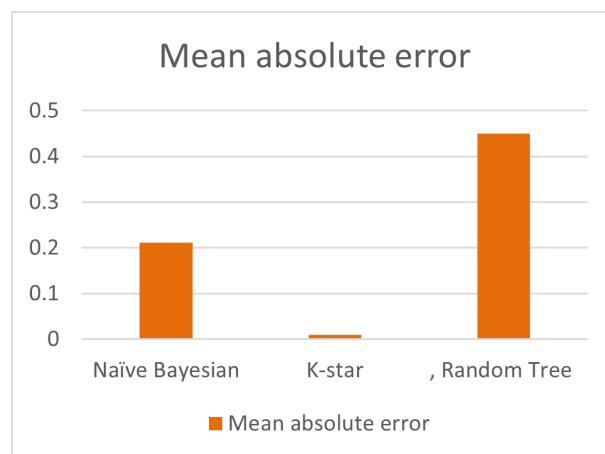Figure 3: Correlation coefficient of mentioned algorithms

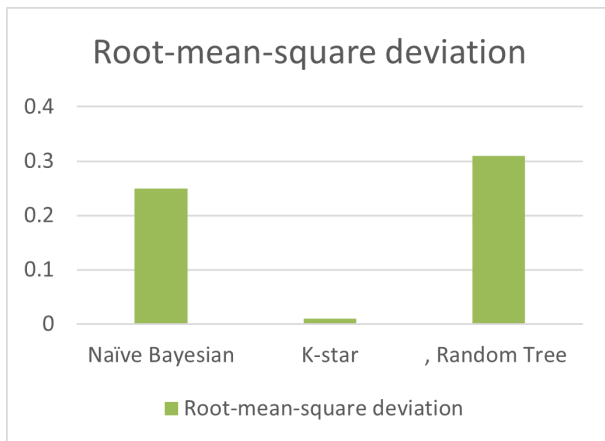Figure 4: Mean Absolute error of mentioned algorithms

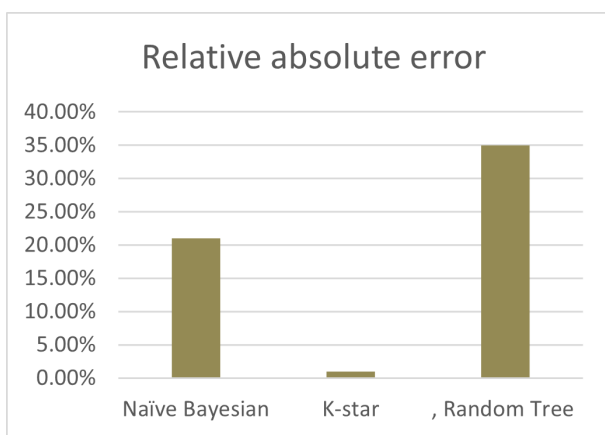Figure 5: Root mean square error of mentioned algorithms

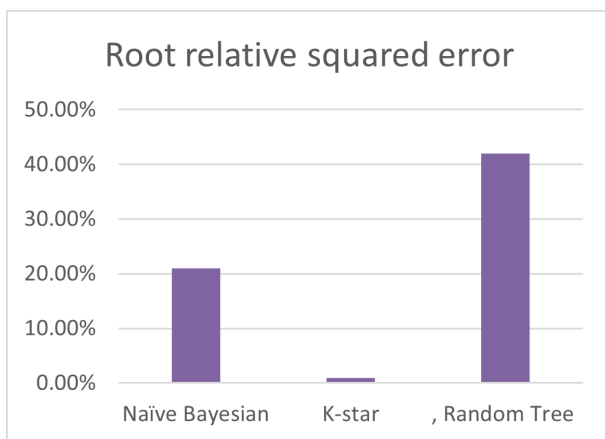Figure 6: The relative absolute error of mentioned algorithms

Figure 7: Root relative squared error of mentioned algorithms

5 Conclusion

Throughout the course of this paper, we provided a method for the accurate identification of WBCs as well as bits based on the categorization, enhancement, and segmentation of images. In order to carry out particular stages of classification, a variety of approaches that already existed were applied. We classified the white blood cell images

using three different classification methods (Naïve Bayes, Random Tree, and the K-star classifier), which resulted in the formation of five distinct groups: basophils, eosinophils, lymphocytes, monocytes, and neutrophils. We then used five distinct evaluation metrics to determine which of the five was the most effective. Based on the evaluation metrics of Mean Absolute Error, Root Means Square Error, Relative Absolute Error, and Root Relative Squared Error, the results showed that the K-star classifier is the best of the mentioned algorithms in this dataset.

References

[1] O.M. Amin Ali, S. Wahhab Kareem, and A.S. Mohammed, *Evaluation of electrocardiogram signals classification using CNN, SVM, and LSTM algorithm: A review*, 8th Int. Engin. Conf. Sustain. Technol. Dev. (IEC) (Erbil, Iraq), IEEE, February 2022, pp. 185–191.

[2] H.Q. Awla, A. Rahman Mirza, and S.W. Kareem, *An automated CAPTCHA for website protection based on user behavioral model*, 8th Int. Engin. Conf. Sustain. Technol. Dev. (IEC) (Erbil, Iraq), IEEE, February 2022, pp. 161–167.

[3] S.K. Bandyopadhyay, *Method for Blood cell segmentation*, J. Global Res. Comput. Sci. **2** (2011), no. 4, 130–135.

[4] S. Banerjee, B.R. Ghosh, S. Giri, and D. Ghosh, *Automated system for detection of white blood cells in human blood sample*, Smart Computing and Informatics, Springer, 2018, pp. 13–20.

[5] R.R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, *WEKA manual for version 3-7-8 2013*, Available at:. Accessed July **21** (2013).

[6] L. Breiman, *Bagging predictors (technical report 421)*, University of California, Berkeley, 1994.

[7] T. Das, *Machine learning algorithms for image classification of hand digits and face recognition dataset*, Machine Learn. **4** (2017), no. 12, 640–649.

[8] T.M. Deserno, *Fundamentals of biomedical image processing*, Biomedical Image Processing, Springer, 2010, pp. 1–51.

[9] A. Gautam and H. Bhadauria, *Classification of white blood cells based on morphological features*, IEEE, 2014, pp. 2363–2368.

[10] R.S. Hawezi, F.S. Khoshaba, and S.W. Kareem, *A comparison of automated classification techniques for image processing in video internet of things*, Comput. Electric. Engin. **101** (2022), 108074.

[11] M.D. Joshi, A.H. Karode, and S.R. Suralkar, *White blood cells segmentation and classification to detect acute leukemia*, Int. J. Emerg. Trends Technol. Comput. Sci. **2** (2013), 147–151.

[12] F. Kamiran and T. Calders, *Data preprocessing techniques for classification without discrimination*, Knowledge Inf. Syst. **33** (2012), no. 1, 1–33.

[13] S.W. Kareem, *An evaluation algorithms for classifying leukocytes images*, IEEE, 2021, pp. 67–72.

[14] M. Kuhn and K. Johnson, *Applied predictive modeling*, vol. 26, Springer, 2013.

[15] S. Loussaief and A. Abdelkrim, *Machine learning framework for image classification*, IEEE, 2016, pp. 58–61.

[16] K. M-Amen, O. Abdullah, A. Amin, Z. Mohamed, B. Hasan, M. Shekha, H. Najmuldeen, F. Rahman, Z. Housein, A. Salih, A. Mohammed, L. Sulaiman, B. Barzingi, D. Mahmood, H. Othman, D. Mohammad, F. Salih, S. Ali, T. Mohamad, K. Mahmood, G. Othman, M. Aali, G. Qader, B. Hussen, F. Awla, S. Kareem, F. Qadir, D. Taher, and A. Salihi, *Cancer incidence in the kurdistan region of Iraq: Results of a seven-year cancer registration in Erbil and Duhok governorates*, Asian Pacific J. Cancer Prevent. **23** (2022), no. 2, 601–615 (en).

[17] D.Y. Mahmood and M.A. Hussein, *Intrusion detection system based on K-star classifier and feature set reduction*, IOSR J. Comput. Engin. **15** (2013), no. 5, 107–12.

[18] H.A. Muhamad, Sh.W. Kareem, and A.S. Mohammed, *A comparative evaluation of deep learning methods in automated classification of white blood cell images*, IEEE, 2022, pp. 205–211.

[19] H.A. Muhamad, S.W. Kareem, and A.S. Mohammed, *A deep learning method for detecting leukemia in real images*, Neuro Quantol. **20** (2022).

[20] S. Nazlibilek, D. Karacor, K.L. Ertürk, G. Sengul, T. Ercan, and F. Aliew, *White blood cells classifications by SURF image matching, PCA and dendrogram*, Biomed. Res. **26** (2015), no. 4, 633–640.

[21] J. Prinyakupt and C. Pluempitiwiriyawej, *Segmentation of white blood cells and comparison of cell morphology by linear and naïve Bayes classifiers*, Biomed. Engin. Online **14** (2015), no. 1, 1–19.

[22] F.L. Quilumba, W.-J. Lee, H. Huang, D.Y. Wang, and R. Szabados, *An overview of AMI data preprocessing to enhance the performance of load forecasting*, IEEE, 2014, pp. 1–7.

[23] R.C. Quinlan, *4.5: Programs for machine learning morgan kaufmann publishers inc*, San Francisco, USA, 1993.

[24] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, Pearson Education India, 2016.

[25] H. Zhou, J. Wu, and J. Zhang, *Digital image processing: part II*, Bookboon, 2010.