ISSN 1870-4069

Systematic Review of the State of the Art Regarding the Identification of Cancer Cells of the Leukemia Type with Digital Image Processing

José de Jesús Moya Mora, Manuel Martín

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

moyamora@me.com, mmartin@cs.buap.mx

Abstract. The present diagnosis is based on a review of the literature of several research projects carried out on the use of tools in the detection of leukemia cancer. This paper presents a discussion of the existing literature to know the current state of knowledge about the identification of leukemia using digital images and areas of opportunity for future work are identified.

Keywords. Acute leukemia, machine learning, deep learning, digital image processing, pattern recognition.

1 Introduction

Machine learning and deep learning tools can provide alternatives regarding how medical problems are solved when giving a diagnosis to a patient. The expansion of the data generated by our systems, the medical literature and the inefficiencies of health care systems will require the use of the power of artificial intelligence tools [4, 31].

The integration of computational tools into medical practice, such as machine learning and deep learning, has begun to be widely used. The U.S. Food and Drug Administration of North America has tested much artificial intelligence-based software since 2017 for medical use [6, 31].

The introduction of digital pathology has revolutionized and provided many opportunities in the improvement of traditional pathology and opened up new research opportunities, such as telemedicine [10, 14].

Recently, the use of digital pathology has allowed the use of machine learning algorithms in the automation of the diagnosis of medical treatments [9, 27]. The challenges facing the use of machine learning, as well as deep learning algorithms in the pathology, are diverse, from the digitalization of cell samples, manual labeling in case of supervised learning, initial and maintenance costs, advanced equipment, technical experience, and ethical consideration. However, the possible opportunities to implement tools in pathology are quite a lot [2, 29].

Leukemia is a hematological neoplasm that confers mortality throughout different ages since this disease includes very early ages as is the case of babies in some cases until appearing in adults. It was estimated that there were around 350,000 thousand new cases in 2012 worldwide [7].

pp. 69–79; rec. 18-08-2019; acc. 15-09-2019

2 Works and Methods

In this article, we review the literature related to the use of machine learning and deep learning in the diagnosis of acute and chronic leukemia, both lymphoid and myeloid. The objective of this review was to understand current trends and limitations in the diagnosis of leukemia. The characteristics included in the designs of the reviewed works, techniques used, properties, as well as identification techniques of the different types of leukemia were analyzed.

Applying the search strategies in the aforementioned databases, 30 publications were found, which were completely reviewed and considered of great interest since they provide information on the methodologies in their studies, as well as the results they obtained.

The evaluation metrics used in the research reviewed in the literature varied, however, all studies gave at least one evaluation metric such as accuracy or sensitivity, which were the most common among the diagnoses reviewed.

2.1 Machine Learning

In the literature reviewed for the identification of leukemia the largest number of studies studied was found lymphoid leukemia as the most studied, which is done through a database of images of the disease, goes through a pre-processing of it to obtain better results as read in the literature which allows to obtain characteristics that better define the cell, later reaches a classification and validation process to mention the result obtained by the methodology used.

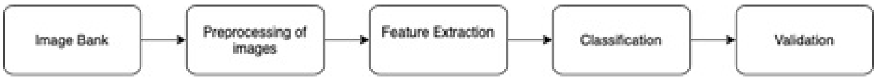

Fig. 1. Block representation of the methodologies used in the literature for the identification of leukemia.

The acquisition of images in most of the diagnoses found uses a database, Image Database (IDB). This digital image library contains two sets of data, the cells of a set are in the original format, meaning that they are not segmented, while the cells of the second set are already segmented.

Image analysis and pattern recognition methods have been widely used in the field of pathological analysis to help specialists study different cell patterns in microscopic images.

There are different types of cells available in the diagnosis made in the blood smear or tissue sections. These include red blood cells, white blood cells, and platelets. They all exist in different types with different characteristics of shape, color, and texture. The diversity of the cells, the existence of staining artifacts and complex scenes, are some of the examples presented to the specialist to determine the type of leukemia that a patient could suffer when reviewing the samples obtained.

Overlapping or clustering of cells could cause segmentation problems, as well as variations in color, contrast, and background due to non-standard staining techniques applied to leukocytes, different thicknesses and lighting conditions [16].

Leukocytes tend to migrate to tissue sections of blood vessels to eliminate harmful agents and begin the healing process [18]. Leukocytes are divided into two main categories according to the structure of the nuclei: polymorphonuclear cells (granular), mononuclear cells (non-granular) Granulocytes have granules in their cytoplasm and are of three types: neutrophils, basophils, eosinophils.

On the other hand, lymphocyte and monocyte cells are non-granular cell types, since they have only one nucleus. It should be mentioned that leukocytes have no color, but they acquire it when they are stained with chemicals to make them visible under the microscope [22], as can be seen in Fig. 2.

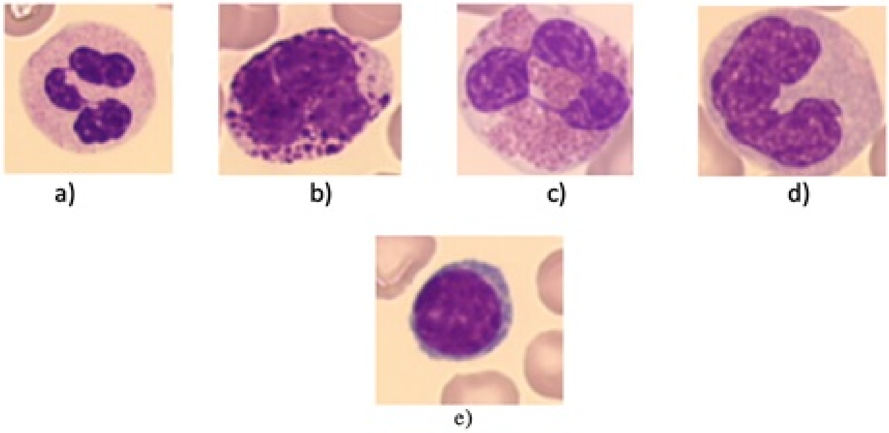

Fig. 2. Representation of white blood cell cells, (a) neutrophil, (b) basophil, (c) eosinophils, (d) monocyte, and (e) lymphocyte [5].

To highlight the colorless leukocytes available in the tissue section or the blood smear, special types of dyes or chemicals are used, and this process is known as staining. The different types of methods used for staining are: Wright staining [33].

Photomicrographs of the tissue section or blood smear may have variations in their color intensity due to the concentration of staining, aging of the staining solution or stained slides, to name a few.

Images of blood smears or tissue sections for clinical and preclinical analysis are widely acquired through bright field microscopy. Image quality is also affected by the use of various types of illuminators such as LED, HBO and XBO [3]. The included works used different approaches to carry out their preprocessing of their databases. The methodology used in the most common cell segmentation algorithm was the recognition of patterns such as geometric or texture, followed by threshold-based methodologies such as Otsu.

Table 1. Different characteristics used to identify leukocytes mentioned in the literature.

Article	Number of images	Feature Extraction Method	Clasificator	Subtypes of cancer cells identified
[17]	98	Area, perimeter and circularity of the lymphocyte nucleus	SVM (Support Vector Machine)	2 classes, normal lymphoid and diseased lymphoid
[18]	220	Gaussian approach, Otsu algorithm, geometric, center and radius of the cells.	MCA (Center point circle algorithm)	2 classes, monocytes and neutrophils, both sick
[19]	106	Segment the white blood cell, HSL color space,	K-means clustering	2 classes basophils and lymphocytes, both sick
[20]	108	Color-based method with search engine of the nearest neighbor with Euclidean distance.	SVM	2 classes, diseased lymphocytes, lymphocyte 1, lymphocyte 2 and lymphocyte 3
[21]	149	Color, geometry and texture analysis	SVM	2 monocyte and neutrophil classes.
[22]	108	Color (green), Otsu?s, morphological operations.	Hough circular transformation	3 classes sick lymphocytes, lymphocyte 1, lymphocyte 2, lymphocyte 3
[23]	300	Color spaces, RGB, HSL, CMYK	K-means clustering	basophilic and monocyte classes both sick.
[24]	288	Characteristics of some blood cell flow cytometry data provided by Beckman-coulter corporation.	SVM	4 normal and sick monocyte classes, healthy and diseased neutrophils.
[25]	150	Geometric and texture characteristics.	SVM y K-PCA	3 monocyte, basophilic and eosinophilic classes all sick.
[26]	160	Characteristics of color space and geometries.	K-means y SVM	4 classes basophils and neutrophils sick and healthy.
[27]	150	Characteristics geometries.	KNN y linear Bayes normal	3 classes diseased lymphocytes, lymphocyte 1, lymphocyte 2 and lymphocyte 3
[28]	256	Statistical and texture characteristics.	EMC-SVM	4 monocyte, lymphocyte, basophil and eosinophil classes all sick.

[29]	166	Statistical characteristics	ELM, RVM	6 monocyte, lymphocyte and basophilic classes, sick and healthy.
[30]	108	Characteristics geometries.	k-nearest neighbor	3 classes, diseased lymphocytes, lymphocyte 1, lymphocyte and lymphocyte 3.
[31]	65	Color and morphological space	SVM	3 classes, basophilic, neutrophils and lymphocyte all sick.
[32]	138	Statistical, texture and geometric characteristics	SVM	2 classes, basophilic, sick and healthy.
[33]	145	Characteristics of color space	Bayesian	2 classes basophilic and eosinophil.
[34]	168	Geometric characteristics and statistics	SVM	4 classes, neutrophils, basophilic, eosinophils and lymphocyte.
[35]	222	Geometric characteristics and statistics	k-nearest neighbor and SVM	4 classes, monocytes, basophilic, neutrophils and eosinophil
[36]	480	Morphological characteristics	Neural network	6 classes, monocytes, lymphocyte, basophilic, healthy and sick
[37]	568	Color spaces	Deep neural network	4 classes monocytes, basophilic, healthy and sick
[38]	345	Geometric, color	Bayesian networks	3 classes lymphocytes sick
[39]	440	Statistics	Neural network	6 classes lymphocytes, monocytes, eosinophil, healthy and sick
[40]	560	Features, geometric, color.	SVM, k-means	4 classes monocytes, basophilic, healthy and sick
[41]	380	Texture, geometric, color	Neural network	4 classes lymphocytes healthy and sick

Most of the works reviewed in table 1 usually segment both the nucleus and the cytoplasm, few studies carry out segmentation in the nucleus of the cell. In the reviewed diagnoses there is not much difference in the evaluation metrics given, from the models in those works that only segmented the nucleus and the cytoplasm to those that only segmented the nucleus. At the time of extracting characteristics, geometry or texture techniques were used.

Some factors that play an important role in the accuracy and classification of leukemia, is how cells are segmented, as well as representations of characteristics used. Characteristic representations must contain useful information, while robust, the background, color, size, location or uneven illumination of the images. Feature extraction representations have been used with different machine learning techniques to classify cells in the blood.

Color and lighting variations can reduce the efficiency of the manual or automated identification system that can lead to a skewed analysis. Therefore, these images should be normalized to minimize variations. This facilitates the best segmentation and, consequently, improves classification accuracy [1]. Several methods of leukocyte classification are proposed in grayscale images that eliminate the need to normalize color but lead to loss of cell color information.

2.2 White Blood Cell Segmentation

In general, the segmentation process involves the removal of leukocytes from the background of the image which then presents a noisy background, where the cell is located, which makes the most vital process difficult. For tissue section images, segmentation becomes more complex due to its complex morphological structures, variable staining, lighting variations, out of focus image components and variability in objects of interest [20].

Cellular segmentation methods can be classified into threshold-based methods [19, 23], methods based on pattern recognition [17], deformable models [12] and metaheuristic-based methods [11, 30].

Threshold methods have also been used as previous steps to locate expected cell locations. Hamghalam [21] used the Otsu threshold method [24] to detect the precise limit of the nuclei in the peripheral blood smear images before applying the active contour method for the detection of the cytoplasm limit [13].

Table 2. Methods of cell segmentation.

Article	Category	Subcategory	Description
17, 18, 19, 20, 21, 22, 23, 34	Thresholding based method	Watershed, Otsu threshold method, growth region	Fast and reliable methods for uniform but consistent images
25, 26, 27, 30, 31, 34, 35, 36	Method based on pattern recognition	SVM, ANN, k-means clustering algorithm, fuzzy c-means	Categorized as unsupervised or supervised methods. Unsupervised methods produce poor results for images that have a complex.
28, 29	Methods metaheuristics	10 point, bold	The segmentation problem is considered an optimization problem and they tried to find the global optimum to segment the objects.

Other variants of threshold-based methods are region-based methods and watershed methods. The performance of the region-based methods is based on the image intensity information. In the region's growth methods, the connected regions of the image are found using seed points and predefined pixel intensity information or border information [28].

In general, seed points are selected manually, which is a major disadvantage of regional cultivation methods [25]. Bread and cabbage [8] used the entropy-based region growth method for leukocyte segmentation. In this work, the regions of interest (of its acronym in English ROI) were located using the special color, and then the entropy of the regions was continuously improved by the expansion of the ROI.

Since leukocytes in microscopic images can be treated as objects, pattern recognition methods can be used to perform segmentation that can be classified as unsupervised or supervised methods.

Unsupervised methods, also known as grouping methods, extract the objects from the data itself. On the contrary, supervised methods use learning-based methods to classify objects. Clustering methods divide the n objects into k groups using the optimization of a criterion function designed for a particular problem. The most used methods in this context are the k-means [26], the diffuse c-means [15].

2.3 Leukocyte Feature Extraction

The outstanding characteristics of the objects to be analyzed are generally used for classification. In microscopic images for the classification of healthy or diseased cells, most of the recent literature focused on texture, size and shape characteristics [26, 32]. In addition to these characteristics, methods based on principal component analysis (PCA) have also been used for feature extraction [8, 26]. In general, the characteristics used in leukocyte analysis can be grouped into geometric characteristics and texture characteristics as shown in table 3.

Table 3. Characteristics used to classify leukocytes.

Article	Category	Features used
17, 18, 19, 20, 21, 22, 23, 24, 3031, 32, 33	Geometric features	Area, radius, convex area, perimeter form factor, concavity, elongation, circularity, symmetry, rectangularity, area ratio, solidity, compactness, concavity
26, 27, 28, 30, 31, 32, 33, 34, 41	Texture characteristics	Variance, standard deviation, correlation, entropy, color, mean, homogeneity, energy, smoothness

For cell discrimination, geometric characteristics play a vital role as they describe the structure and size of leukocytes. The area and perimeter of the leukocytes are the characteristics used to represent the size of the cells, while the shape characteristics can be grouped into characteristics based on regions and boundaries. To extract the characteristics, the cell image is converted into a binary image where the pixels of the cell are represented with a non-zero value.

The characteristics of shape and size are briefly analyzed below:

- Area: the area of the cell is represented by the total number of non-zero pixels within the cell limit.
- Area ratio: the area ratio is defined as the ratio between the number of pixels in the cytoplasm and the number of pixels in the nucleus.
- Convex area: in some cases, the convex hull is calculated, and its area is called the number of pixels within its limit.
- Symmetry: symmetry represents the difference between lines that are perpendicular to the axis greater than the cell limit in both directions.
- Concavity: Concavity is defined as the extent to which the actual limit of a cell is within each chord between non-adjacent limit points.
- Elongation: elongation is the measure of the relationship between the maximum distance and the minimum distance from the center of gravity to the core limit.
- Number of lobes: nuclear polymorph cells have a different number of nuclei (lobes) in their cytoplasm that can be used as one of the most prominent features for cell sorting.
- Orientation: the angle between the x axis and the major axis of the cell is known as orientation.
- Circularity: the circularity of the cell is defined by the ratio of the perimeter of the narrowest bounding circle to the perimeter of the cell.
- Rectangularity: the rectangularity of the core is represented as the ratio between the perimeter of the narrowest bounding rectangle and the perimeter of the core.
- Perimeter: calculated by measuring the sum of the distances between successive limit pixels.

The different properties mentioned were used by different researchers for leukocyte differentiation [1, 5, 20, 23, 33]. These characteristics also play an important role in reducing different noisy elements present in the images. Piuri and Scotti [15, 21] used 13 different characteristics for the identification of leukocytes in blood smear images. But most researchers used texture features along with geometric features to increase the accuracy of the classifier.

3 Discussion

Leukemia is a major hematological malignancy, with high prevalence and incidence, leukemia diagnosis is facing multiple changes.

Leukocytes classification is the complex process for computer systems as compared to human observer which has motivated the researchers to study this field in the perspective of artificial intelligence.

Automated leukocytes classification can help the pathologists in the disease identification and drug development. Although, a significant amount of work has been done in this field in past two decades, there are still some challenges which lead to lower accuracy of cell classification.

Majority of the included studies in this review used supervised learning algorithms. A drawback of using supervised learning in pathology is the need to label samples which is time-consuming and might introduce errors. A solution to that would be to use unsupervised learning methodologies, in which the patterns are determined by the data itself.

One of the major problems in the analysis of tissue section images is its complex morphological structure. It is difficult to differentiate leukocytes from other histological structures such as cells of hair follicles and basal cells of epidermis, red blood cells and other artifacts produced during processing and staining. Another limitation is the variability in the shape of leukocytes mainly due to the plane of sectioning and age of leukocyte in the inflamed tissues, which impose criticality in cell identification.

4 Conclusions

Automated diagnostic studies used a variety of segmentation methodologies of both the nucleus and cytoplasm. The most used method was the method based on pattern recognition, with diffuse c-averages as the most used methodology. Fuzzy c-mean has proven to be more accurate than the grouping of k-means according to the literature investigated. All studies have extracted geometric and texture characteristics. The included studies represented many limitations in research, both in machine learning and in deep learning. These limitations include issues such as sample size, generalization and prospective analysis.

The models presented in this work will reach high accuracy, commonly more than 90%. This is a very common result in the field of machine learning and deep learning research in the field of research that was on cancer cells. This can be attractive; However, it can pose different problems. First, the models presented in this review are generally based on a small sample size and, in many studies, the data come from a single center, which raises the question of how machine learning models are proposed. Therefore, these studies must use more robust databases that will need records and large digital libraries with the ability to avoid the limitation of overfitting.

References

1. Abbas, K., Banks, J., Chandran, V., Tomeo-Reyes, I., Nguyen, K.: Classification of White Blood Cell Types from Microscope Images:Techniques and Challenges, pp. 17–25 (2018)
2. Acs, B., Rimm, D.L.: Not Just Digital Pathology, Intelligent Digital Pathology. JAMA Oncology 4(3), 403–404 (03 2018)
3. Adjouadi, M., Zong, N., Ayala, M.: Multidimensional pattern recognition and classification of white blood cells using support vector machines. Particle & Particle Systems Characterization - PART PART SYST CHARACT 22 (2005)
4. Beam, A.L., Kohane, I.S.: Big Data and Machine Learning in Health Care (2018)

5. Bhavnani, L., Jaliya, U., Joshi, M.: Segmentation and counting of wbcs and rbcs from microscopic blood sample images. International Journal of Image, Graphics and Signal Processing 8, 32–40 (2016)
6. Digital Health Criteria: Food and Drug Administration. https://www.fda.gov/MedicalDevices/Digital Health/ucm575766.htm (2019)
7. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F.: Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. International Journal of Cancer 136(5), E359–E386 (2015)
8. Ghaisani, F., Wasito, I., Faturrahman, M., Mufidah, R.: Deep belief networks and bayesian networks for prognosis of acute lymphoblastic leukemia. In: ICACS '17: Proceedings of the International Conference on Algorithms, Computing and Systems. pp. 102–106 (2017)
9. Ghaznavi, F., Evans, A., Madabhushi, A., Feldman, M.: Digital imaging in pathology: Whole-slide imaging and beyond. Annual Review of Pathology: Mechanisms of Disease 8(1), 331–359 (2013)
10. Golden, J.: Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. JAMA 318, 2184 (2017)
11. Jiang, K., Liao, Q.M., Dai, S.Y.: A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. vol. 5, pp. 2820 – 2825 Vol.5 (2003)
12. Joshi, M.M.D., Karode, A.H., Suralkar, P.S.R.: White blood cells segmentation and classification to detect acute leukemia (2013)
13. Kim, K., Jeon, J., Choi, W., Kim, P., Ho, Y.S.: Automatic cell classification in human's peripheral blood images based on morphological image processing. vol. 2256, pp. 225–236 (2001)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60 – 88 (2017)
15. Macawile, M.J., Quinones, V., Ballado, A., Cruz, J., Caya, M.V.: White blood cell classification and counting using convolutional neural network. pp. 259–263 (2018)
16. Ramoser, H., Laurain, V., Bischof, H., Ecker, R.: Leukocyte segmentation and classification in blood-smear images. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. pp. 3371–3374 (2005)
17. Ravikumar, S.: Image segmentation and classification of white blood cells with the extreme learning machine and the fast relevance vector machine. Artificial Cells, Nanomedicine, and Biotechnology 44(3), 985–989 (2016)
18. Ruberto, C.D., Loddo, A., Putzu, L.: A leukocytes count system from blood smear images segmentation and counting of white blood cells based on learning by sampling (2016)
19. Sajjad, M., Khan, S., Jan, Z., Muhammad, K., Moon, H., Kwak, J.T., Rho, S., Baik, S.W., Mehmood, I.: Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities. IEEE Access 5, 3475–3489 (2017)
20. Salah, H., Muhsen, I., Salama, M., Owaidah, T., Hashmi, S.: Machine learning applications in the diagnosis of leukemia: Current trends and future directions. International Journal of Laboratory Hematology 41 (2019)
21. Sanei, S., Lee, T.: Cell recognition based on pca and bayesian classification. In 4th International Symposium, ICA (2003)
22. Sarrafzadeh, O., Rabbani, H., Talebi, A., Banaem, H.: Selection of the best features for leukocytes classification in blood smear microscopic images. vol. 9041 (2014)
23. Scotti, F.: Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. pp. 96–101 (2005)
24. Segura, M., Rivero, R., Suárez, V., Machado, M., Martínez, E., Otero, A., Abraham, C., Hernández Ramírez, P.: Inmunofenotipaje en el diagnóstico de síndromes linfo y mielo-

proliferativos. Revista Cubana de Hematologia, Inmunologia y Hemoterapia 16, 198–205 (2000)

25. Shafique, S., Tehsin, S.: Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technology in Cancer Research & Treatment 17 (2018)
26. Teman, C.J., Wilson, A.R., Perkins, S.L., Hickman, K., Prchal, J.T., Salama, M.E.: Quantification of fibrosis and osteosclerosis in myeloproliferative neoplasms: A computer-assisted image study. Leukemia Research 34(7), 871 – 876 (2010)
27. Theera-Umpon, N.: White blood cell segmentation and classification in microscopic bone marrow images. vol. 3614, pp. 787–796 (2005)
28. Tizhoosh, H., Pantanowitz, L.: Artificial intelligence and digital pathology: Challenges and opportunities. Journal of Pathology Informatics 9(1), 38 (2018)
29. Tomasz, M., Leszek, M.: Analysis of features for blood cell recognition. pp. 42–45 (2004)
30. Topol, E.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25 (2019)
31. Tran, T., Vununu, C., Atoev, S., Lee, S.H., Kwon, K.R.: Leukemia blood cell image classification using convolutional neural network. International journal of computer theory and engineering 10, 54–58 (2018)
32. Vogado, L.H., Veras, R.M., Araujo, F.H., Silva, R.R., Aires, K.R.: Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification. Engineering Applications of Artificial Intelligence 72, 415 – 422 (2018)
33. Zhang, C., Xiao, X., Li, X., Chen, Y.J., Zhen, W., Chang, J., Zheng, C., Liu, Z.: White blood cell segmentation by color-space-based k-means clustering. Sensors (Basel, Switzerland) 14, 16128–16147 (2014)