CONESCAPANHONDURAS2025paper135.pdf

Institute of Electrical and Electronics Engineers (IEEE)

Document Details

Submission ID

trn:oid:::14348:477749691

Submission Date

Jul 31, 2025, 10:09 PM CST

Download Date

Aug 12, 2025, 6:30 PM CST

File Name

CONESCAPANHONDURAS2025paper135.pdf

File Size

3.2 MB

6 Pages

3,383 Words

20,206 Characters

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

29 Not Cited or Quoted   13%
Matches with neither in-text citation nor quotation marks

4   Missing Quotations 1%
Matches that are still very similar to source material

3   Missing Citation 2%
Matches that have quotation marks, but no in-text citation

0   Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13%   🌐   Internet sources

13%   📖   Publications

0%   👤   Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

29 Not Cited or Quoted **13%**
Matches with neither in-text citation nor quotation marks

4 Missing Quotations **1%**
Matches that are still very similar to source material

3 Missing Citation **2%**
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted **0%**
Matches with in-text citation present, but no quotation marks

Top Sources

13%	🌐	Internet sources
13%	📖	Publications
0%	👤	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet		
www.mdpi.com			**2%**

2	Internet		
www.coursehero.com			**2%**

3	Publication		
Sumair Aziz, Muhammad Bilal, Muhammad Umar Khan, Fatima Amjad. "Deep Lea...			**<1%**

4	Publication		
Aimin Zhou, Jianyong Sun, Qingfu Zhang. "An Estimation of Distribution Algorith...			**<1%**

5	Internet		
ir.lib.nchu.edu.tw			**<1%**

6	Internet		
www.medicalsystems.it			**<1%**

7	Internet		
www.hindawi.com			**<1%**

8	Publication		
Juliana Melendrez-Ruiz, Laurence Dujourdy, Isabelle Goisbault, Jean-Christophe C...			**<1%**

9	Publication		
Eduardo Dávila-Meza, Eduardo Bayro-Corrochano. "Quaternion and Split Quatern...			**<1%**

10	Internet		
yushen-liu.github.io			**<1%**

11   Publication

Sweeney, K. T., T. E. Ward, and S. F. McLoone. "Artifact Removal in Physiological Si...   <1%

12   Internet

fastercapital.com   <1%

13   Internet

centrodeconocimiento.ccb.org.co   <1%

14   Publication

Minh Long Hoang, Guido Matrella, Paolo Ciampolini. "Artificial Intelligence Imple...   <1%

15   Internet

dl.icdst.org   <1%

16   Internet

www.thaiscience.info   <1%

17   Internet

c.coek.info   <1%

18   Internet

export.arxiv.org   <1%

19   Internet

www.businessresearchinsights.com   <1%

20   Publication

Imran Ahmed, Eulalia Balestrieri, Ioan Tudosa, Francesco Lamonaca. "Segmentati...   <1%

21   Internet

pdfs.semanticscholar.org   <1%

22   Internet

pure.au.dk   <1%

23   Publication

Richard R. Khan. "The AI Glossary - Demystifying 101 Essential Artificial Intelligen...   <1%

24   Internet

sibgrapi.sid.inpe.br   <1%

25 **Internet**

www.fastercapital.com **<1%**

26 **Publication**

"Surveillance, Prevention, and Control of Infectious Diseases", Springer Science a... **<1%**

27 **Publication**

Liqun Lin, Weixing Wang, Bolin Chen. "Leukocyte recognition with convolutional ... **<1%**

28 **Internet**

escholarship.org **<1%**

29 **Publication**

Wang, Xingqiao. "Customization of Large Language Models for Causal Inference ... **<1%**

30 **Publication**

Meilu Zhu, Jing Liao, Jun Liu, Yixuan Yuan. "FedOSS: Federated Open Set Recogniti... **<1%**

Monocyte and Lymphocyte Classifier from Blood Smear Digital Images Using K-Means Clustering

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Abstract—The analysis and classification of white blood cells (leukocytes) is essential for diagnosing numerous pathologies, as these cells represent the first line of defense in the immune system. However, manual leukocyte differential counting is a time-consuming process that requires trained personnel specialized in microscopic identification and is prone to human error, especially when a single specialist must process a large volume of blood smears. This work proposes an automatic classification algorithm based on K-means clustering, specifically focused on classifying agranulocytes into lymphocytes and monocytes, which are crucial for diagnosing infections such as mononucleosis, tuberculosis, and hepatitis. A key aspect of this proposal is its focus on local technological development, prioritizing independence from high-cost foreign commercial solutions. Preliminary results demonstrate the feasibility of the method, achieving an overall accuracy of 77%, with a precision of 93% and a recall of 77%. These findings support the creation of local know-how for transparent, customizable, and accessible biomedical solutions.

Index Terms—Agranulocytes, Automatic classification, HSV, Image segmentation, K-means, Lymphocyte, Monocyte, Morphological operation, White blood cells.

I. INTRODUCCIÓN

Due to their role as key immune system agents, white blood cells (leukocytes) are crucial biomarkers for diagnosing a wide range of medical conditions. Their functions include detecting and eliminating infectious and tumoral agents, as well as coordinating the inflammatory response. An imbalance in their number or morphology can compromise the organism's ability to fight infections. They represent approximately 1% of total blood volume and are classified into five main types, each with specialized functions: neutrophils, which phagocytize bacteria and increase during acute infections (sometimes presenting a "left shift"); lymphocytes, which participate in adaptive immunity; monocytes, which mature into macrophages and eliminate cellular debris; eosinophils, which combat parasites and mediate allergic reactions; and basophils, which release chemical mediators during inflammatory processes. Morphological alterations such as changes in the nucleus or chromatin may be indicative of leukemias, myeloproliferative disorders, or severe infections, making the accurate classification of leukocytes a key diagnostic tool [1].

To assess the quantity and morphology of leukocytes, a peripheral blood smear is traditionally performed. This involves spreading a drop of blood onto a microscope slide, allowing it to air-dry, and staining it with differential dyes. The most commonly used staining method is Wright–Giemsa, which combines eosin and methylene blue to highlight the nucleus and cytoplasm of various cell types. Under an optical microscope, this technique allows clear distinction between white blood cells, red blood cells, and platelets, as well as the identification of morphological abnormalities—such as changes in nuclear size, chromatin patterns, or the presence of cytoplasmic inclusions—that may indicate leukemia, severe infections, or other hematological disorders [2]–[4].

Leukocyte classification is fundamental for interpreting inflammatory markers and predicting patient health status. However, various factors such as age, sex, immune condition, and environmental influences can affect leukocyte counts. Therefore, it is necessary to consider specific reference ranges and conduct repeated measurements to ensure reliable classification [5]. From a technical perspective, manual smear analysis provides detailed morphological evaluation, but it is time-consuming, requires skilled personnel, and is subject to high interobserver variability. In response to these limitations, clinical practice has adopted automated hematology analyzers capable of classifying leukocytes into five or three groups depending on the model, based on impedance cytometry and laser light scattering principles [6].

However, while these devices provide rapid and standardized counts, they have limited capacity for morphological analysis, hindering the accurate identification of all leukocyte types and detection of rare abnormalities. Additionally, their high cost restricts their use in primary care settings, which prevents them from being considered a fully efficient solution [7].

Several studies have addressed this issue by exploring digital image processing techniques as a feasible alternative to conventional analyzers. For instance, methods combining color-based segmentation, K-means clustering, and mathematical morphology have been proposed for the automatic extraction of leukocyte nuclei from RGB images, reducing reliance on expensive equipment and minimizing manual intervention [8]. These approaches demonstrate the potential of computer

vision to support hematological analysis in resource-limited environments.

In this context, the present study proposes an alternative approach based on digital processing of blood smear images to support automated leukocyte classification. The goal is to develop a reliable, low-cost diagnostic support tool suitable for resource-limited environments, preserving diagnostic accuracy and morphological analysis without compromising clinical effectiveness [9], [10].

II. Methodology

In this work, a four-stage methodology is proposed for the classification of lymphocytes and monocytes: preprocessing, segmentation, morphological operations, and cell classification.

A. Dataset

For the development and evaluation of the automatic leukocyte classification system, a widely recognized public database was used: *A dataset of microscopic peripheral blood cell images for development of automatic recognition systems (Matek et al., 2019)*. This dataset contains 18,365 microscopic images of peripheral blood cells in RGB format, with a resolution of 360×360 pixels, captured using conventional optical microscopy and manually labeled by hematology experts. Although the complete dataset includes 16 different cell types, only the classes corresponding to lymphocytes and monocytes were selected in this study, as they are of specific interest for validating the classification algorithm. The annotations are organized in a .csv file, allowing for direct integration into automated training and evaluation processes.

B. Image Preprocessing

The process begins with the reading of RGB-format images from an input folder. To facilitate subsequent segmentation, the images are converted to the HSV (Hue, Saturation, Value) color space (Figure 1) [11], as this model separates color information (hue and saturation) from intensity (value). This transformation enhances perceptual contrast in cell images, where nuclei and cytoplasm tend to differ more clearly in hue and saturation than in their original RGB values. Additionally, in blood smear images—where nuclei and cytoplasm are often more distinguishable by hue and saturation than by pure color intensity—the use of the HSV space significantly improves perceptual contrast, thereby contributing to more effective segmentation.

The hue (H) and saturation (S) channels are used to construct a feature matrix for unsupervised segmentation. The value (V) channel is reserved for identifying structures based on their brightness level, since nuclei typically exhibit lower intensity than the cytoplasm or the background.

C. Segmentation Using K-Means

To segment the nuclear and cytoplasmic cell structures, the K-means clustering algorithm is applied. This algorithm uses the HS matrix, where each row represents a pixel with its
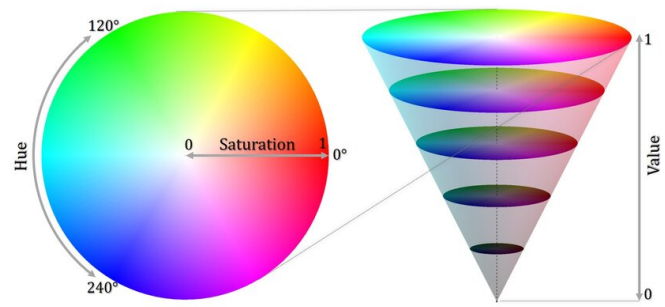

Fig. 1. HSV color model [12].

corresponding values, and groups the pixels into three clusters (K=3), using 5 replicates ('Replicates', 5) and a maximum of 200 iterations ('MaxIter', 200) to ensure stable convergence. The clustering results are reshaped using reshape to visualize the segmentation through imshow and label2rgb, assigning different colors to each cluster.

After classifying the pixels using the clustering results, the cluster corresponding to the nucleus and the one corresponding to the cytoplasm are identified by analyzing the brightness (V) channel, assuming the nucleus is the darkest region. Using sort(means), the clusters are ordered from darkest to brightest, assuming the nucleus corresponds to the darkest (order(1)) and the cytoplasm to the next (order(2)). This allows the creation of two initial masks: one for the nucleus (segmented == idx_nucleus) and another for the cytoplasm (segmented == idx_cytoplasm), which contain only the pixels corresponding to the nucleus and cytoplasm, respectively, for subsequent morphological analysis.

D. Classification Algorithm

The obtained masks exhibit typical imperfections of automatic segmentation, so morphological operations are applied for their refinement. For the nucleus, hole filling (imfill), erosion with a square structuring element (strel('square',6)), and removal of small objects using bwareaopen are performed, considering a minimum area threshold (<500 pixels). For the cytoplasm, similar steps are performed but adapted to its morphology. The mask is filled, merged with the nucleus mask through a logical OR operation (since both form the cell body), and a mild dilation is applied using a circular structuring element (strel('disk',1)). Regions connected to the image border are removed (imclearborder), as well as small structures (<6000 pixels). A final fill ensures a continuous and clean mask. From the processed masks, the area occupied by the nucleus and cytoplasm is calculated, expressed in number of pixels. These values serve as the basis for classifying the cell as a lymphocyte or monocyte, according to previously established morphometric criteria. For example, a lymphocyte is characterized by a proportionally large nucleus relative to a small cytoplasm, whereas a monocyte exhibits the opposite proportion. Finally, as shown in Figure 2, the contours of the nucleus (in red) and cytoplasm (in green) are overlaid on the original image, providing an intuitive visualization of the
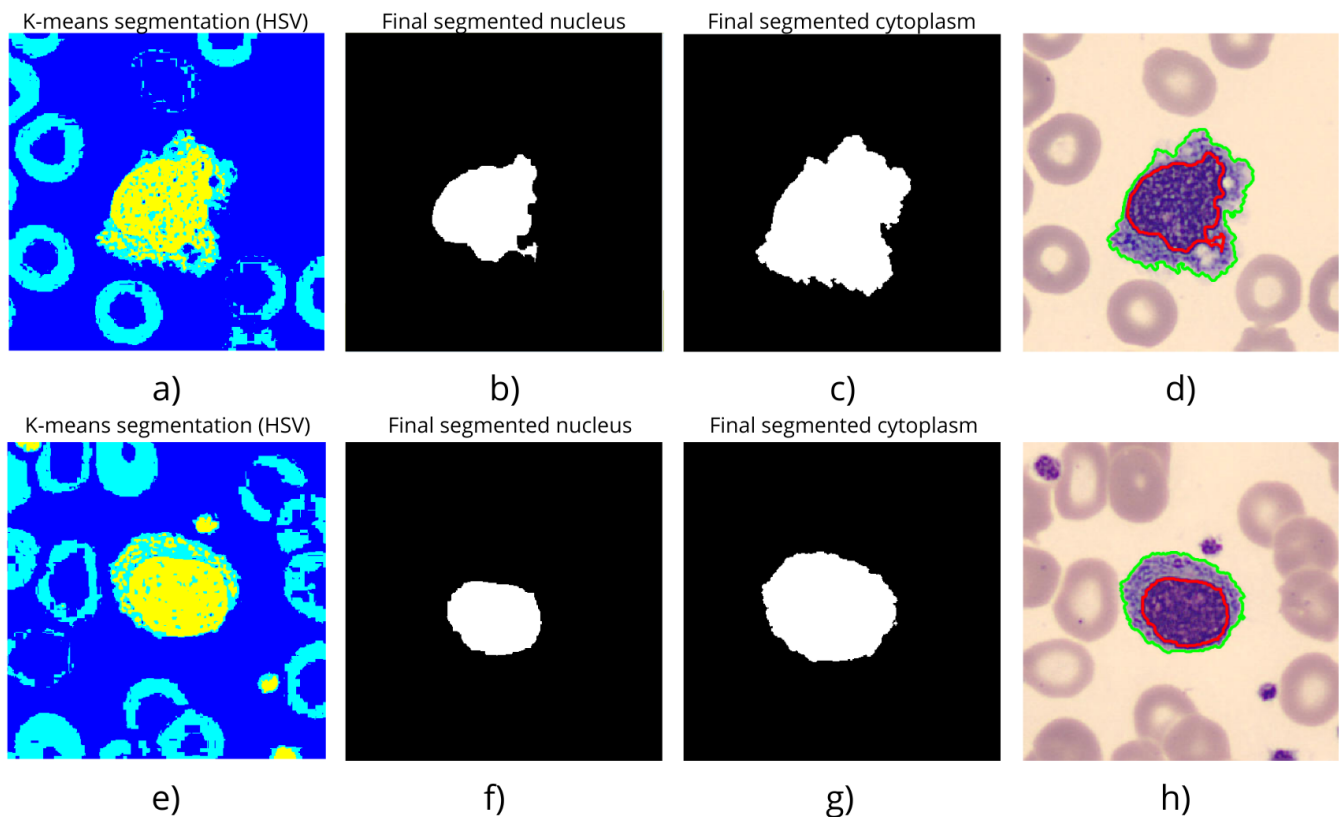
Fig. 2. a) RGB image of a monocyte converted to HSV space and segmented using k-means. b) Final monocyte nucleus. c) Final monocyte cytoplasm. d) Area calculation. e) RGB image of a lymphocyte converted to HSV space and segmented using k-means. f) Final lymphocyte nucleus. g) Final lymphocyte cytoplasm. h) Area calculation.
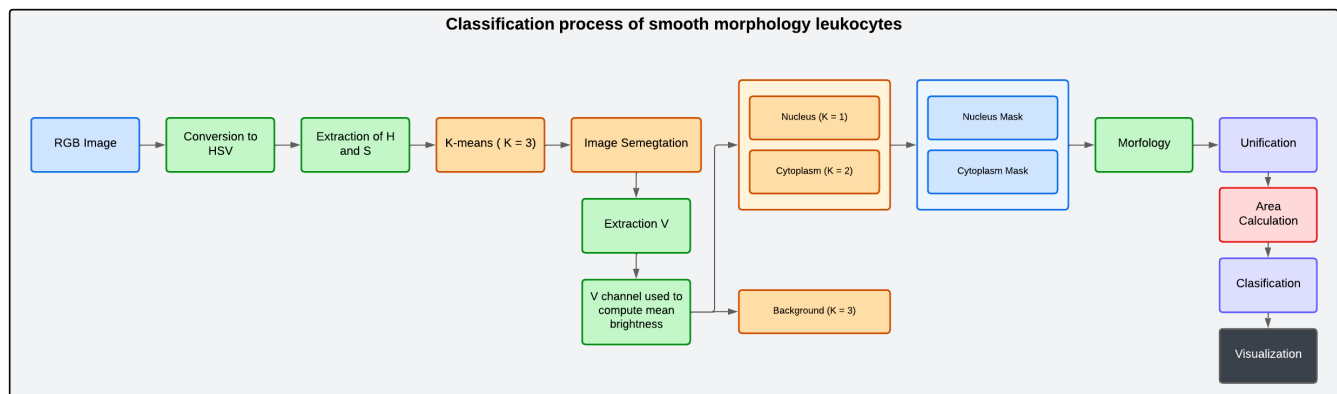

Fig. 3. General processing scheme using k-means.

result. Additionally, the area values and the assigned class for each cell are displayed, allowing for visual and quantitative validation of the procedure. The complete classification process is shown in Figure 3; likewise, the technical specifications of the equipment used are presented in Table I.

TABLE I
SPECIFICATIONS OF THE SYSTEM USED

Componente	Specification
Operating System	Windows 11 Home
Processor (CPU)	Core™ i5-1135G7 @2.40GHz
RAM Memory	8GB DDR4
Graphics Processing Unit (GPU)	Intel® Iris® Xe Graphics 128MB
Storage	SSD 240 GB
Software de desarrollo	Matlab2021a

III. RESULTS

A. Classification based on cell area

To enable the automatic classification of leukocytes into monocytes and lymphocytes, criteria were established based on specific ranges for nucleus and cytoplasm areas, as illustrated in Figure 4. These thresholds were defined following a detailed statistical analysis of a representative set of previously segmented images. The study revealed consistent morphological differences between the two cell types, particularly in the relative size of their components. Monocytes exhibit nuclei and cytoplasmic regions that are significantly larger in size compared to those of lymphocytes, which provides a useful discriminative feature for their classification.

Based on these observations, the following area ranges were determined: a cell is classified as a monocyte if its nucleus area is between 6,180 and 11,590 pixels, and its cytoplasm area is between 11,865 and 48,460 pixels. Conversely, a cell is considered a lymphocyte if its nucleus area ranges from 4,100 to 8,220 pixels, and its cytoplasm area ranges from 6,500 to 16,600 pixels. These criteria, grounded in quantifiable morphological characteristics, enable reliable and automated classification of cell samples.

Table II summarizes the results obtained by applying these criteria, showing the number of correct and incorrect classifications for each leukocyte type, and supporting the preliminary validation of the proposed method.
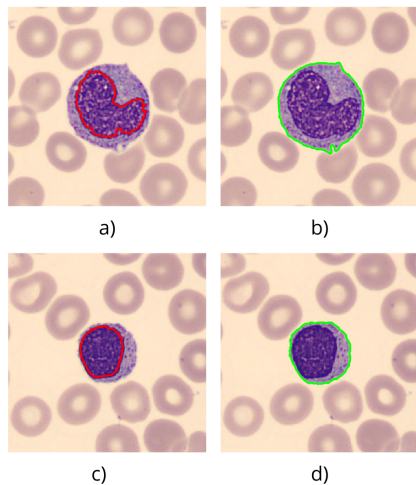

Fig. 4. a) Monocyte nucleus b) Monocyte cytoplasm c) Lymphocyte nucleus d) Lymphocyte cytoplasm

TABLE II
THE RESULTS OF TESTING BY LEUKOCYTES FEATURES SET

Leukocytes	Samples	Correct	Incorrect
Linfocito	1214	873	341
Monocito	1430	1166	264
Total	2644	2039	605

B. Performance of the sorting algorithm

In this study, an unsupervised classification algorithm was employed to differentiate between two types of leukocytes: monocytes and lymphocytes. The model was validated using a confusion matrix and standard evaluation metrics such as accuracy, precision, and recall. These metrics enable assessment of the algorithm's performance from multiple perspectives, thereby providing a comprehensive evaluation.

The confusion matrix, shown in Figure 5, illustrates the relationship between the model's predictions and the actual classes of the samples. In this representation, rows correspond to true labels, while columns represent the predicted labels. The numerical values indicate the number of samples classified in each prediction–ground truth combination. Table III summarizes the key performance values derived from this matrix.
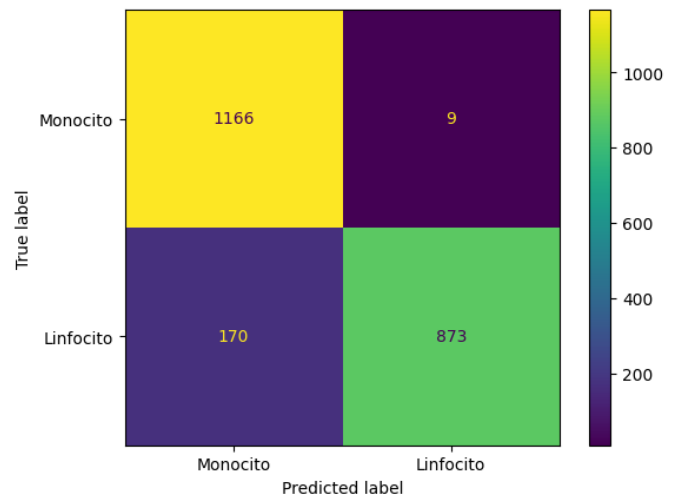

Fig. 5. Confusion Matrix

- 1166: True positives (TP) for the monocyte class. These samples were correctly identified by the model as monocytes.
- 9: False positives (FP) for the lymphocyte class. These correspond to actual lymphocyte samples that were misclassified as monocytes.
- 170: False negatives (FN) for the lymphocyte class. These are lymphocyte samples that the model failed to correctly identify and were classified as monocytes.
- 873: True negatives (TN) for the lymphocyte class. Samples correctly classified as lymphocytes.

From this matrix, the following metrics were calculated:

Accuracy measures the proportion of correct predictions (true positives and true negatives) out of the total evaluated samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad [13]$$

In this case, the accuracy reached 0.77 (77%). This indicates the model correctly classified 77% of samples, representing acceptable performance considering the dataset exhibits some

degree of class imbalance. However, accuracy alone proves insufficient to fully assess model performance, particularly in cases where one class may be overrepresented.

Precision measures the proportion of positive predictions that truly belong to the positive class. This metric is particularly valuable for evaluating the reliability of the model's positive classifications:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [14]$$

A weighted precision of 0.93 (93%) was obtained. When analyzing this metric by class, the precision for lymphocytes was 99%, indicating a very low false positive rate for this class. In contrast, the precision for monocytes was 87%, demonstrating that the model also performs well in correctly identifying these cells. These results show that when the model predicts a sample belongs to a particular class, it is correct in the vast majority of cases.

Recall (also known as sensitivity) measures the proportion of actual class samples that were correctly classified by the model. This metric is particularly useful when minimizing false negatives is important:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [15]$$

The weighted overall recall value was 0.77 (77%). Specifically, 82% was obtained for the monocyte class and 72% for the lymphocyte class. This indicates that most true monocyte samples are correctly identified, but greater difficulty exists with lymphocytes, where approximately 28% of the samples were misclassified (false negatives). This result can be attributed to the complexity of cellular morphology or class imbalance in the training dataset.

TABLE III
SUMMARY OF CLASSIFICATION METRICS

Class	Precision (%)	Recall (%)	Accuracy (%)
Monocyte	87	82	77
Lymphocyte	99	72	77
Overall	**93**	**77**	**77**

The results obtained indicate that the model has a high ability to make reliable positive predictions (high precision), although its capability to identify all positive samples of each class (recall) is moderate, especially for lymphocytes. This type of behavior is common in models trained with imbalanced data, where better identification of the dominant class is observed.

One of the main limitations identified in the classification of lymphocytes and monocytes lies in the considerable morphological variability of the cells, which may be influenced by factors such as differences in staining technique, variations in image acquisition quality, as well as the presence of artifacts or unwanted elements in the visual field. This variability introduces an additional level of complexity in the segmentation and morphological analysis stages, especially

when using conventional methods and a limited dataset. Likewise, the visual similarity between certain cell types poses a significant challenge for classification algorithms, highlighting the need to employ more robust models and larger, better-curated databases.

During processing, cases have been detected in which some cells appear partially represented, either because they are cropped at the image margins or exhibit structural deformations, which hinders accurate segmentation and negatively affects feature extraction. Additionally, as shown in Figure 6, when red blood cells are close to the cell nucleus, the system tends to confuse these elements with the cytoplasm, thereby reducing the overall accuracy of the classification process.

Cell 1
Nucleus area: 5785 px² | Cell body area: 15827 px²

Cell 2
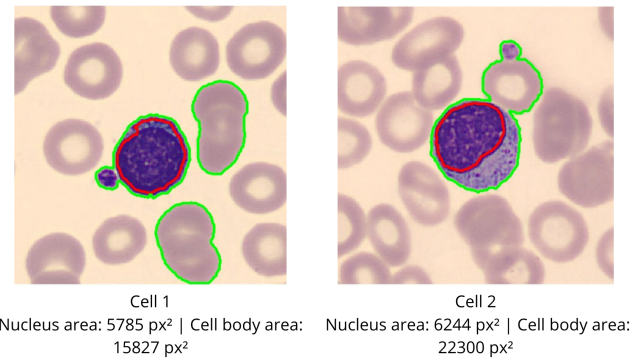Nucleus area: 6244 px² | Cell body area: 22300 px²

Fig. 6. Confusion of the cytoplasm with red blood cells due to their proximity.

As future research directions, the system could incorporate strategies such as data balancing, as well as the implementation of more advanced artificial intelligence models, including deep neural networks specialized in morphological analysis. These improvements would increase the classifier's sensitivity without sacrificing precision, especially when expanding coverage to a larger number of cell classes and increasing dataset diversity. This approach would significantly contribute to enhancing the system's performance in real clinical contexts, where cellular heterogeneity and variable sample quality are critical factors to consider.

IV. CONCLUSIONS

This study demonstrated the feasibility of using the K-means clustering algorithm in the HSV color space for the automatic classification of monocytes and lymphocytes from digital microscopic blood smear images, employing digital image processing techniques implemented in MATLAB. The experimental results obtained from 2,644 samples showed an overall accuracy of 77%, with a high precision of 93% and a recall of 77%, indicating reliable positive predictions, although with a lower sensitivity for lymphocytes (72%).

A key limitation was the appearance of false positives in images in which the proximity between agranulocytes and erythrocytes hindered accurate segmentation. Future research could address this issue by incorporating deep learning techniques, larger and more balanced datasets, or additional features such as texture and shape to enhance model robustness.

Despite these limitations, this work lay the groundwork for the development of accessible and cost-effective hematological tools. The proposed solution, validated on 2,644 samples, demonstrates that algorithms such as K-means clustering can enable more efficient health monitoring.

As an economical and scalable approach, the developed tool may serve as a complementary resource for healthcare professionals, facilitating the early detection of hematologic disorders and contributing to timely treatment interventions.

REFERENCES

[1] M. M. WINTROBE, "Diagnostic significance of changes in leukocytes," *Bulletin of the New York Academy of Medicine*, vol. 51, no. 4, pp. 357–361, Jun 1975, disponible en PMC: PMC1911375. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1911375/

[2] S. Wick, "Human physiology and anatomy: Blood cell histology," http://www.unomaha.edu/hpa/blood.html, 1997, [Accessed: Dec. 23, 2013].

[3] C. Cuevas, Ed., *Hematología: Fundamentos y aplicaciones clínicas*, 2nd ed. México: McGraw-Hill Education, 2015, coordinador.

[4] K. B. Poulsen, "Morfología y función de los componentes celulares," in *Biología Celular y Molecular*, K. B. Poulsen, Ed. McGraw-Hill Interamericana, ch. 5, p. No disponible, organización celular, Membrana celular, Núcleo, Citoplasma.

[5] P. P. Chmielewski, B. Strzelec, E. Y. Kalafi, C. Town, and S. K. Dhillon, "Elevated leukocyte count as a harbinger of systemic inflammation, disease progression, and poor prognosis: a review," *Folia Morphologica*, vol. 77, no. 2, pp. 171–178, 2018.

[6] B. K. Timby and N. E. Smith, *Introductory Medical-Surgical Nursing plus Live Advice Online Student Tutoring Service*. Philadelphia, PA, USA: Lippincott Williams and Wilkins, 2006.

[7] A. Kratz, S.-h. Lee, G. Zini, J. A. Riedl, M. Hur, S. Machin, and the International Council for Standardization in Haematology, "Digital morphology analyzers in hematology: Icsh review and recommendations," *International Journal of Laboratory Hematology*, vol. 41, no. 4, pp. 437–447, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijlh.13042

[8] A. Gautam and H. Bhadauria, "White blood nucleus extraction using k-mean clustering and mathematical morphing," in *2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*, 2014, pp. 549–554.

[9] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*. Elsevier Saunders, 2006.

[10] H. Kim, M. Hur, G. d'Onofrio, and G. Zini, "Real-world application of digital morphology analyzers: Practical issues and challenges in clinical laboratories," *Diagnostics*, vol. 15, no. 6, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/15/6/677

[11] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing using Matlab*, 4th ed. Upper Saddle River, NJ: Pearson Education, 2018, authorized adaptation from the United States edition, entitled *Digital Image Processing*, Fourth Edition.

[12] E. Davila-Meza, "Hsv color model. this figure was made using matlab software," 2025, available: https://www.researchgate.net/figure/HSV-color-model-This-figure-was-made-using-MATLAB-software_fig1_373704120. Accessed: Jul. 9, 2025.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[15] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.