RESEARCH ARTICLE

VOL. 43 | NO. 2 | **MAY - AUGUST 2022** | PP 39-52

Revista Mexicana de Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.43.2.3

E-LOCATION ID: 1254

Comparison of Accuracy of Color Spaces in Cell Features Classification in Images of Leukemia types ALL and MM

Comparación de Precisión de Espacios de Color en la Clasificación de Características de Células en Imágenes de Leucemia tipos ALL y MM

Cinthia Espinoza-Del Angel, *Aurora Femat-Diaz*

Universidad Autónoma de Querétaro

ABSTRACT

This study presents a methodology for identifying the color space that provides the best performance in an image processing application. When measurements are performed without selecting the appropriate color model, the accuracy of the results is considerably altered. It is significant in computation, mainly when a diagnostic is based on stained cell microscopy images. This work shows how the proper selection of the color model provides better characterization in two types of cancer, acute lymphoid leukemia, and multiple myeloma. The methodology uses images from a public database. First, the nuclei are segmented, and then statistical moments are calculated for class identification. After, a principal component analysis is performed to reduce the extracted features and identify the most significant ones. At last, the predictive model is evaluated using the k-nearest neighbor algorithm and a confusion matrix. For the images used, the results showed that the CIE L*a*b color space best characterized the analyzed cancer types with an average accuracy of 95.52%. With an accuracy of 91.81%, RGB and CMY spaces followed. HSI and HSV spaces had an accuracy of 87.86% and 89.39%, respectively, and the worst performer was grayscale with an accuracy of 55.56%.

KEYWORDS: PCA, Statistical moments, Color spaces, Leukemia images

RESUMEN

Este estudio presenta una metodología para identificar el espacio de color que proporciona el mejor rendimiento en una aplicación de procesamiento de imágenes. Cuando las mediciones se realizan sin seleccionar el modelo de color adecuado, la precisión de los resultados se altera considerablemente. Esto es significativo en el procesamiento, principalmente cuando el diagnóstico se basa en imágenes de microscopía de células teñidas. Este trabajo muestra cómo la selección adecuada del modelo de color proporciona una mejor caracterización en dos tipos de cáncer, la leucemia linfoide aguda y el mieloma múltiple. La metodología utiliza imágenes de una base de datos pública. Primero, se segmentan los núcleos y luego se calculan los momentos estadísticos para la identificación de clases. Posteriormente, se realiza un análisis de componentes principales para reducir las características extraídas e identificar las más significativas. Por último, el modelo predictivo se evalúa utilizando el algoritmo k-vecinos más cercanos y una matriz de confusión. Para las imágenes utilizadas, los resultados mostraron que el espacio de color CIE L*a*b caracterizó mejor los tipos de cáncer analizados con una precisión promedio del 95,52%. Con una precisión del 91,81%, siguieron los espacios RGB y CMY. Los espacios HSI y HSV tuvieron una precisión del 87,86% y el 89,39%, respectivamente, y el peor desempeño fue la escala de grises con una precisión del 55,56%.

PALABRAS CLAVE: PCA, Momentos estadísticos, Espacios de color, Imágenes de leucemia

Corresponding author

TO: Aurora Femat-Diaz
INSTITUTION: Universidad Autónoma de Querétaro
ADDRESS: Facultad de Ingeniería, Universidad Autónoma de Querétaro, Cerro de las Campanas S/N, Col. Las Campanas, Centro, C. P. 76010, Santiago de Querétaro, Querétaro, México
CORREO ELECTRÓNICO: afemat@uaq.mx

Received:

11 March 2022

Accepted:

9 May 2022

INTRODUCTION

Leukemia is a blood disease distinguished by the abnormal production of white blood cells [1]. Its diagnosis uses a blood smear where the presence of myeloblasts or lymphoblasts is determined [2] [3]. This examination is usually a time-consuming manual process and requires microscopist expertise [4] [5] [6]. Recently, image processing techniques with machine learning have been used, which integrate image processing and segmentation, feature extraction and selection, and a classification algorithm [7] [8]. The most critical steps are segmentation and selection of significant features [9] [10] [11].

During microscopic analysis, cells are stained to provide visibility and contrast [12]. The segmentation stage separates the cells from the rest of the image from the acquired color. Among the techniques that have been applied to segment are K means [13] [14] [15] [16] [17] [18], Fuzzy c-means [19], Triangle thresholding [20] [21] [22], and Otsu thresholding [9] [23] [24] [25] [26], which are usually accompanied by the Watershed algorithm to divide adjacent or overlapping cells [27] [28].

From the nucleus or cytoplasm, parameters are calculated that help to identify cancer types. Several features can be analyzed, including geometric, statistical, and texture [21] [26] [29]. The number of parameters used should be limited using a reduction algorithm to improve the efficiency of classification model [30] [31].

Some methods to decrease the number of characteristics are Univariate feature selection (k-Best) [32], Social Spider Optimization Algorithm (SSOA) [33], Genetic Algorithm (GA) [25], Statistically Enhanced Salp Swarm Algorithm (SESSA) [34], Linear Discriminant Analysis (LDA) [35] and Principal Component Analysis (PCA) [35] [36] [37]. The latter is a statistical technique that reduces the dimension of a data set and generates a new set of uncorrelated variables. These are called principal components (PC), and their relationship preserves the maximum variation from the original data set [38] [39].

On the other hand, color models are used to define the way to represent the tones mathematically. The color spaces RGB [9] [26], CMYK [16] [20] [35], HSI [23] [40], HSV [19] [24], and CIE L*a*b [14] [41], and grayscale [25] [42] [43] have been used for this type of application. Studies have identified that RGB is not ideal for the segmentation of these cells, while HSI, HSV, and CMY perform better [44] [45]. In a group of images where the capture brightness effect varies, the use of HSV space may be the most appropriate because it separates the image intensity from the color information [17] [19].

In feature extraction, statistical and color properties have been obtained from various spaces such as, RGB [21] [30], HSV [13] [19] [46], HSI [35] and CIE L*a*b [14]. Of these, RGB and HSV spaces are the most widely used, but the use of these representations has not been justified [14] [21] [33] [47]. Although statistical and color features are an important source of information, no studies have yet been performed to compare the accuracy of color space using these characteristics in cell sorting with staining.

This paper proposes to use a principal component analysis with statistical descriptors as input variables to determine the color space that best represents the information of a set of images. This process is analyzed by using the k nearest neighbors (kNN) algorithm and a confusion matrix to determine the accuracy of the predictive model. The objective of the study is to propose a tool to image processing methodologies to identify the model that best represents the content of the region of interest (ROI). In particular, it is applied to identify two types of cancer, acute lymphoid leukemia (ALL) and multiple myeloma (MM).

MATERIAL AND METHODS

Image Dataset Definition

The set of images used in this work is a collection of microscopic bone marrow images of patients diagnosed with B-lineage acute lymphoid leukemia and

multiple myeloma, published in The Cancer Imaging Archive (TCIA) by Gupta, A. and Gupta, R. [48]. They have a resolution of 2560x1920 pixels, were captured using a Nikon Eclipse-200 microscope at 1000x magnification, and slides were stained with Jenner-Giemsa stain. This study used 60 samples, 30 ALL and 30 MM.

Nucleus Segmentation of Leukemia Cells

Based on studies by Jagadev and Virani [17], Mirmohammadi *et al.* [19], and Rahman and Hasan [21], the image segmentation was performed using the HSV space, considering that it separates the intensity of the image from the color information, and the images of the database do not have uniform brightness. For the segmentation process, first, the intensity of the images was adjusted to a range of 0.1 to 0.7 using the value channel (V) to decrease the capture luminance effect [19]. The ROI was defined with a threshold value in the Hue (H) and Saturation (S) components. After a binary segmentation, the holes were filled using a morphological closure with disk shape structure element of radius 4 [49]. The watershed algorithm was used to the resulting image to separate the overlapping nuclei. Subsequently, objects with an area of fewer than 17500 pixels and elements with an eccentricity greater than 0.80 and solidity less than 0.65 were removed [14] [21] [50]. The ROI was established from the original images using the binary mask. A total of 484 blast cell nuclei, 168 ALL and 316 MM, were then extracted. The segmentation results for two sample images are visualized in Figure 1.

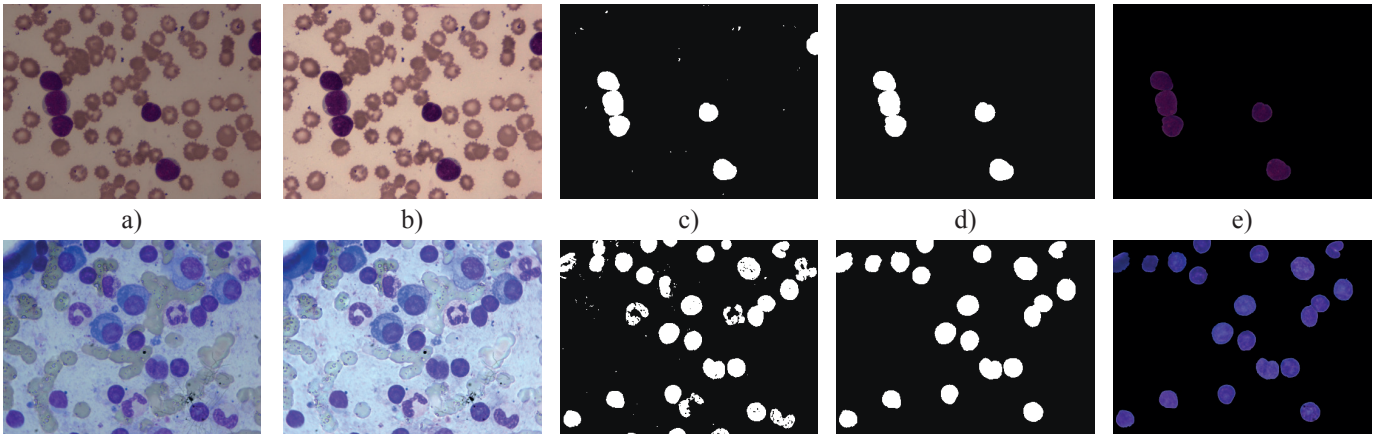


FIGURE 1. Steps of Leukemic Cell Segmentation. ALL original image is shown in a). ALL Image enhancement in b). ALL binary mask in c). ALL filtering and watershed segmentation in d). ALL segmentation result in e). MM original image in f). MM Image enhancement in g). MM binary mask in h). MM filtering and watershed segmentation in i). MM segmentation result j).

Feature Extraction

Mean, *variance*, *standard deviation*, *skewness*, *kurtosis*, *entropy*, and *energy*, were calculated for each nucleus from an image for each color space channel and grayscale. Equations 1 to 7 show the definition of each one of these parameters, where Z represents the intensity as a random variable, $p(Z_i)$, i= 0, 1, 2, ..., L-1 is the probability of occurrence of the value Z_i and L is the number of different possible values [51]. For each leukemia cell, a total of 21 features were obtained in each color space and 7 for grayscale.

$$mean = \sum_{i=0}^{L-1} Z_i \, p(Z_i) \tag{1}$$

$$Variance = \sum_{i=0}^{L-1} (Z_i - m)^n \, p(Z_i) \tag{2}$$

$$SD = \sqrt{\sum_{i=0}^{L-1}(Z_i - m)^n \, p(Z_i)} \quad (3)$$

$$Skewness = \sum_{i=0}^{L-1}(Z_i - m)^3 p(Z_i) \quad (4)$$

$$Kurtosis = \sum_{i=0}^{L-1}[(Z_i - m)^4 p(Z_i)]^{-3} \quad (5)$$

$$Entropy = -\sum_{i=0}^{L-1} p(Z_i)\log_2 p(zi) \quad (6)$$

$$Energy = \sum_{i=0}^{L-1} p^2(Z_i) \quad (7)$$

Feature Selection

The most significant features of each color space were identified using principal component analysis. First, the number of statistical descriptors in the dataset was reduced using the "statistics" and "FactoRMine" libraries of Rstudio software [52] [53]. For this, we first checked for a low partial correlation value between each pair of features, using the Kaiser-Meyer-Olkin (KMO) coefficient. If for any of these pairs, *mean-variance*, *mean-standard deviation*, *mean-skewness*, etc., the KMO coefficient is less than 0.5, the value indicates that it is not appropriate to use PCA in that model. Then Bartlett's test of sphericity (BTS) with a significance level of $p<0.05$ is used to estimate the correlation between variables.

PCA for each color model was performed based on an initial data table represented by a matrix of 484 rows, containing 168 observations of ALL and 316 MM type cells. The measurement result for each statistical descriptor is shown by columns (21 characteristics for the RGB, CMY, HSV, HSI, and CIE L *a*b color spaces and 7 for grayscale). Table 1 shows an example of this for the RGB model components.

TABLE 1. Data matrix for RGB for an example image. *Mean* (M), *variance* (V), *standard deviation* (SD), *Skewness* (S), *energy* (Er) and *entropy* (Et). The initial of RGB channels was added to each descriptor.

Sample	MR	VR	SDR	SR	KR	ErR	EtR	MG	…	EtG	MB	…	EtB
ALL_1_1	85.26	266.19	16.32	-0.22	0.71	0.02	0.74	42.10	…	0.59	103.07	…	0.62
ALL_1_2	85.09	253.95	15.94	-0.48	0.07	0.02	0.74	42.70	…	0.59	105.67	…	0.63
ALL_1_3	88.99	69.71	8.35	0.31	0.13	0.03	0.63	42.63	…	0.56	106.09	…	0.55
⋮	…	…	…	…	…	…	…	…	…	…	…	…	⋮
⋮	…	…	…	…	…	…	…	…	…	…	…	…	⋮
MM_30_10	143.32	99.41	9.97	1.33	3.03	0.03	0.65	65.77	…	0.65	120.03	…	0.52
MM_30_11	97.95	337.36	18.37	0.93	2.23	0.02	0.77	43.93	…	0.60	106.36	…	0.61
MM_30_12	122.91	138.44	11.77	0.42	0.08	0.02	0.70	57.04	…	0.67	115.60	…	0.54

Each table column is standardized to an average of 0 and a standard deviation of 1 using Equation 8, where X_j is the value to be standardized, X_{js} represents the standardized value and, μ_x and σ_x are the average and the standard deviation of the column.

$$X_{js} = \frac{X_j - \mu_x}{\sigma_x} \quad (8)$$

A covariance matrix was calculated using the standardized values for each table to estimate the correlation and dependence between variables. Equation 9 was used to evaluate covariances between each pair of characteristics. Where σ_{jk} is the covariance between the two variables, X_j and X_k represent the standardized value of variables j and k, μ_j and μ_k are the column averages of variables j and k, and n is the total data per

column. The correlation coefficient of the covariances was determined by Equation 10. It is obtained by dividing the covariance by the standard deviations of X_j and X_k represented by σ_j and σ_k.

$$\sigma_{jk} = \frac{1}{n}\sum_{i=1}^{n}(X_j - \mu_j)(X_k - \mu_k) \tag{9}$$

$$\sigma_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_j - \mu_j)(X_k - \mu_k)}{\sigma_j \sigma_k} \tag{10}$$

The covariance matrix C is represented as in Equation 11, where $Cov_{(i,\ j)}$ is the covariance between the elements in row i and column j. This matrix is decomposed into its eigenvalues and eigenvectors to determine the principal components. By solving Equation 12, the eigenvalues λ_k are obtained and for each of them, its eigenvector V_k is determined using Equation 13, where I is the identity matrix. The eigenvectors correspond to the principal components, and the eigenvalues define the magnitude of the variance of the new set of variables. Finally, the eigenvectors are sorted in descending order to select the components that retain at least 80% of the information from the original data set.

$$C = \begin{bmatrix} \sigma_{1,1} & Cov_{1,2} & \dots & Cov_{1,m} \\ Cov_{2,1} & \sigma_{2,2} & \dots & Cov_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Cov_{n,1} & Cov_{n,2} & \dots & \sigma_{n,m} \end{bmatrix} \tag{11}$$

$$det(\lambda I - C) = 0 \tag{12}$$

$$(\lambda_k \mathrm{I} - \mathrm{C}) * V_k = 0 \tag{13}$$

ALL and MM Classification

The ALL and MM cancer types were classified using the principal components or new variables. For this purpose, the k nearest neighbor algorithm was employed using the Rstudio software [54]. Data was randomly divided into two sets. 80% of them were used as training data and the remaining 20% as test data.

In the training phase, the kNN algorithm stores the set of input variables to establish a relationship between them and the conditions to be classified, calculating a distance between the rows of training data and the test set data. It was determined as a function of the Euclidean distance (ED) using Equation 14, where A and B represent the principal component vectors A= (x_1, x_2, x_3, x_4, ... x_m), B= (y_1, y_2, y_3, y_4, ... y_m), and m, the dimensionality of the feature space.

$$ED\ (A,B) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{14}$$

The resulting vector was ordered from smallest to largest so that the smallest distance is considered the k nearest neighbor.

Subsequently, a number for k was defined to determine the nearest neighbors to include in the voting process using Equation 15. N represents the total data in the training set. The operation resulted in a k = 17.

$$k = \sqrt{N} \tag{15}$$

Finally, to determine the performance of the kNN classifier, the confusion matrix was applied. The metric provided is the accuracy and is defined by Equation 16. Where TP is the true positives, FN is the false negatives, FP indicates the number of false positives, and TN is the number of true negatives.

$$Accuracy\ = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

Statistical Analysis

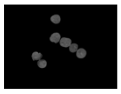
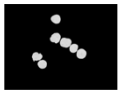
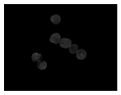
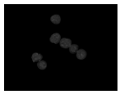
Lastly, differences of accuracy between color and grayscale spaces were analyzed using a one-way analysis of variance (ANOVA) and Tukey's posthoc test. The sample size was calculated using statistical power analysis. A significance level of 0.05 was established,

with a power of 0.8 and an effect size of 0.25. The study determined a sample size of 35. The normality of the residuals was assessed by performing the Shapiro Wilks test, and Bartlett's test demonstrated the homogeneity of variances. For each analysis, was used a significance level of $p<0.05$.

RESULTS AND DISCUSSION

This section shows the results of the comparison between color spaces and grayscale. Table 2 shows an example of channel decomposition and grayscale conversion of an example image.

TABLE 2. Channel division of color representations. Channel 1, Channel 2 and Channel 3 correspond to the color band of each color space.

Color System	Channel 1	Channel 2	Channel 3
Grayscale			
RGB			
CMY			
HSI			
CIE L*a*b			
HSV			

As presented in Table 1, KMO coefficient was calculated; in all cases, the value was greater than 0.5. Bartlett's test was performed considering the parameters of all images as input, presenting for each color space a significant p-value less than $\alpha = 0.05$. Thus, the application of the PCA is adequate. The results of these tests are shown in Table 3.

TABLE 3. Results for Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) and Bartlett's test of sphericity (BTS). Chi-square (χ^2), degrees of freedom (*df*), p value (*p*), statistic (stat) and grayscale (GS).

Test	Stat	GS	RGB	CMY	HSI	L*a*b	HSV
KMO		0.61	0.72	0.72	0.68	0.71	0.62
BTS	χ^2	5779	20621	20621	20728	20133	18346
	df	21	210	210	210	210	210
	p	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

The summary of the PCA is shown in Table 4. The principal components were selected according to the total variance method, retaining at least 80% of the information of the original set of variables. In RGB, CMY, CIE L*a*b, and HSI spaces, 4 PCs were taken; in HSV, 5 PCs; and in grayscale, 2 PCs.

TABLE 4. Principal component analysis summary. The percentage of cumulative variance (Cum. Var) retained by the PC's are highlighted in blue. Principal component (PC), eigenvalue (λ), variance (Var).

Color System	PC	λ	Var (%)	Cum. Var (%)
Grayscale	1	3.49	49.9	49.9
	2	2.73	38.9	**88.8**
RGB	1	8.02	38.2	38.2
	2	5.95	28.3	66.5
	3	1.72	8.21	74.7
	4	1.46	6.94	**81.7**
CMY	1	8.02	38.2	38.2
	2	5.95	28.3	66.5
	3	1.72	8.21	74.7
	4	1.46	6.94	**81.7**
HSI	1	6.42	30.6	30.6
	2	5.18	24.7	55.2
	3	4.00	19	74.3
	4	1.85	8.82	**83.1**
CIE	1	7.53	35.9	35.9
	2	6.00	28.6	64.5
	3	2.04	9.71	74.2
	4	1.49	7.11	81.3
HSV	1	5.82	27.7	27.7
	2	4.51	21.5	49.2
	3	3.11	14.8	64.0
	4	1.96	9.36	73.3
	5	1.83	8.71	**82.0**

The results show that grayscale retains more information in its first two components than the color spaces (see Table 4 data highlighted in blue). The CIE L*a*b space has the highest information loss of 81.3%.

A two-dimensional PCA space was projected for each color model considering the first two components with the highest contribution, as shown in Figure 2. The upper and right coordinates (abscissa and ordinate axis for PC1 and PC2 loadings) show the degree of contribution to the principal components. In these graphs, the black vectors called loadings represent the statistical descriptors. Charge vectors range is from -1 to 1. Charges close to |1| indicate that the variable strongly influences the principal component; those close to 0 denote a weak influence.

A coefficient greater than |0.5| is considered significant to define a PC. For example, for Figure 1a corresponding to grayscale PCA, the *mean* (M), *skewness* (S), *kurtosis* (K), and *energy* (Er) contribute most to PC1. *Variance* (V), *standard deviation* (SD), *energy* (Er) and *entropy* (Et) contribute most to PC2.

The angle between loadings is their correlation. Vectors with equal directions are positively correlated, and those with opposite directions are negatively correlated. If they present an angle of 90°, there is no correlation. The vertical and horizontal axes show the percentage of variability explained by the principal components. The points on the graph are the channel measurements of the color models. Cancer types are grouped in concentration ellipses, ALL cells in blue and MM cells in orange.

As shown in Figure 2, vector loadings differ between color representations, yet it is possible to establish a reliable prediction model because the loadings are significant. Although the clusters of the two cancers overlap at some points in the two-dimensional PCA space, a clear separation is visualized between the ALL and MM descriptors (Figure 2a to Figure 2f). The precision results when evaluating the predictive model using kNN are shown in Table 5. Each of the 35 samples was studied using a different random data set.

TABLE 5. kNN Accuracy results for each color model. Data are in percentage. Grayscale (GS). Average (μ).

No.	GS	RGB	CMY	HSI	L*a*b	HSV
1	86.6	89.7	89.7	87.6	95.9	92.8
2	87.6	91.8	91.8	93.8	100.0	92.8
3	86.6	90.7	90.7	84.5	93.8	86.6
4	85.6	89.7	89.7	86.6	95.9	86.6
5	88.7	95.9	95.9	93.8	96.9	93.8
6	84.5	93.8	93.8	88.7	97.9	89.7
7	84.5	88.7	88.7	80.4	96.9	86.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
34	78.4	92.8	92.8	89.7	95.9	87.6
35	88.7	91.8	91.8	87.6	93.8	85.6
μ	84.7	91.8	91.8	87.9	95.5	89.4

A one-way ANOVA was then applied to compare the performance of each color model statistically. First, the normality of the residuals was tested using the Shapiro-Wilks test; the p-value = 0.0947 showed that they did follow normal behavior. Bartlett's test determined compliance with homogeneity of variance between treatments with a p-value of 0.2215, suggesting no evidence of statistically significant variation between color representations for ANOVA.

The variance analysis revealed a significant difference between the color models with a p-value less than 0.05 (See Table 6).

TABLE 6. One-way ANOVA results. Sum of squares (SS), mean square (MS).

Source	Df	SS	MS	F value	p-value
Model	5	2436	487.2	54.12	<2e-16
Error	204	1836	9.0		

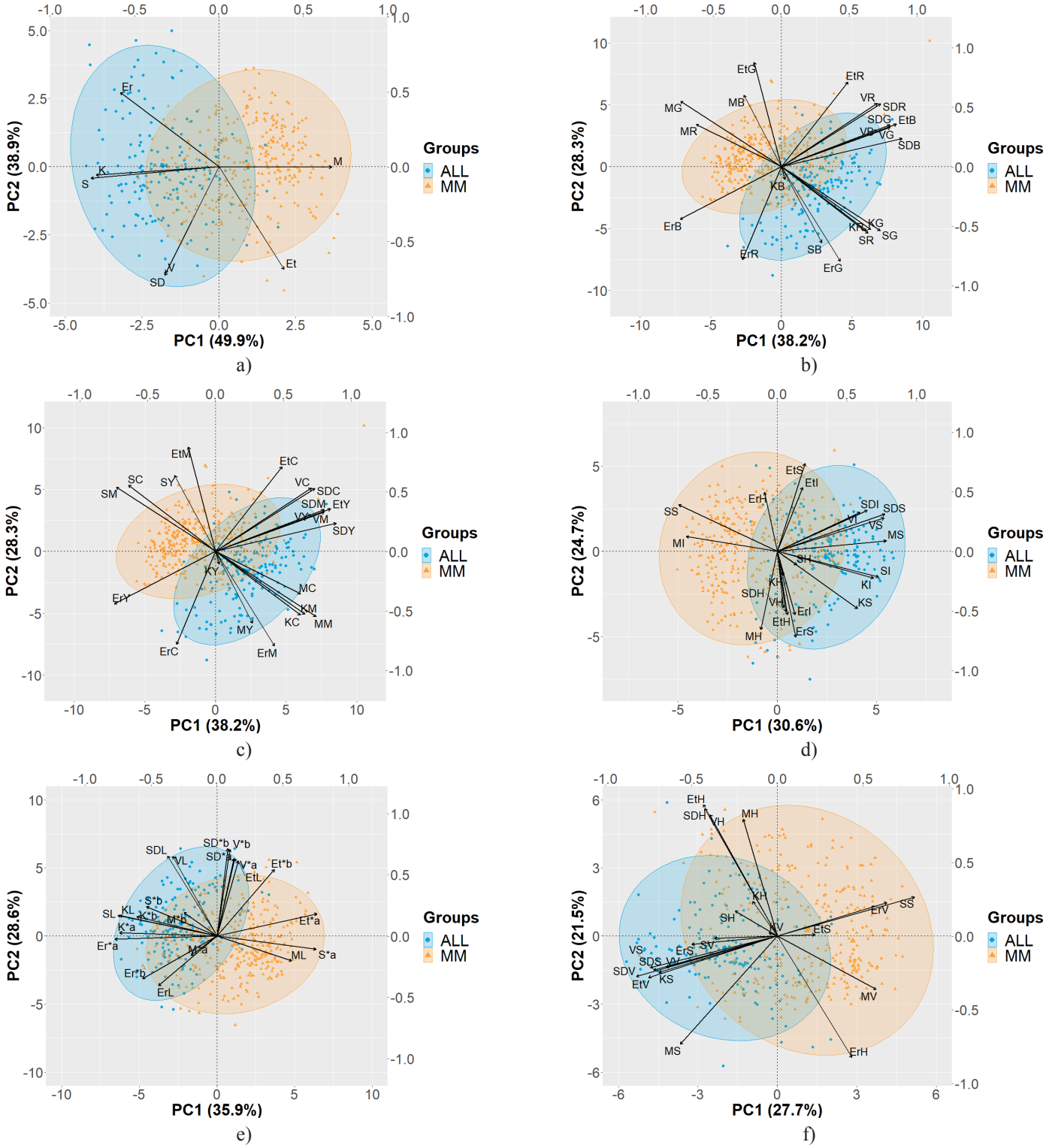


FIGURE 2. **PCA biplot for components 1 and 2. Grayscale is shown in a), RGB in b), CMY in c), HSI in d), CIE L*a*b in e), and HSV in f). Ellipses represent a concentration of the scores for each group set with 95% confidence boundaries. *Mean* (M), *variance* (V), *standard deviation* (SD), *Skewness* (S), *energy* (Er) and *entropy* (Et).**

In Figure 3, the distribution of color representations exhibits a normal distribution; outliers are minimal. The whiskers in the box represent the boundaries of the precision samples drawn for each group. The red dot indicates the mean accuracy of color representation. In grayscale was 84.7; in RGB, 91.8; in CMY, 91.8; in HSI, 87.9; in HSV, 89.4; and in CIE L*a*b, 95.5. The predictive model given by the characteristics of the CIE L*a*b color space obtained the highest accuracy, followed by the RGB and CMY spaces. Grayscale was the worst performer.

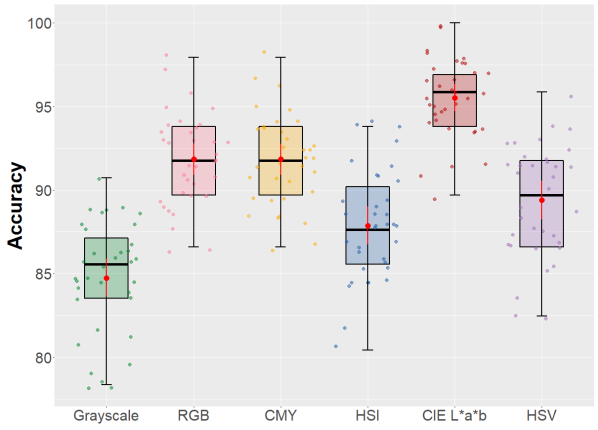


FIGURE 3. Anova results for each color model.

The pairwise comparisons of means of the color models obtained by Tukey's test can be seen in Figure 4.

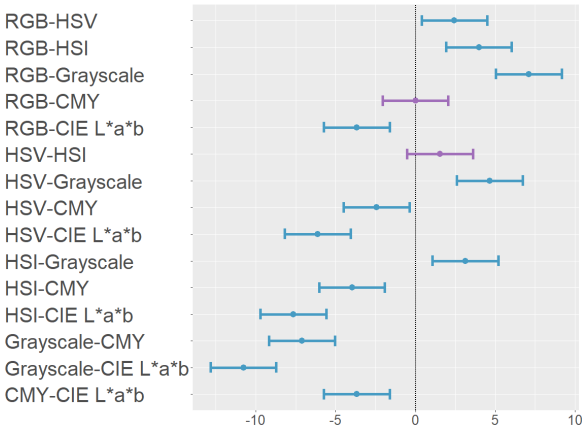


FIGURE 4. Tukey post hoc test pairwise comparison plot. Extended lines in blue color show statistically significant differences between the pairs of means, and extended lines in purple indicate that there is no statistical difference between the means.

In the graph, the extended lines show the 95% confidence intervals. Those crossing the 0 points indicate that there is no statistically significant difference between the pairs of means. The analysis revealed that there is no inequality between RGB and CMY spaces (p-value = 1.00); and HSI and HSV spaces (p-value = 0.27408).

CONCLUSIONS

The use of machine learning and image processing methods play an important role in image analysis for prognosis and early detection of blood cancer. Research for leukemia image classification techniques has used color characteristics from RGB and HSV spaces [21] [30] [13] [19] [46], but other color models are rarely used.

This article presents a methodology to compare the accuracy of different color models to represent the characteristics of leukemia cells. Of the color spaces analyzed, the CIE L*a*b best described the two cancer types, ALL and MM, using color moments with an average accuracy of 95.52%.

Compared to reference articles, the accuracy obtained in this study was superior to that of [36], which used RGB and grayscale space color and texture features. The PCA and the KNN and SVM classifiers were used, obtaining an accuracy of 91.45% and 92.63%, respectively.

Likewise, a better performance was obtained than [30], which used the color characteristics of the RGB space and compared six classifiers; KNN (80.7%), tree classifier (75.8%), ANN (83.5%), logistic regression (82.4%), random forest (81.0%) and SVM (73.6%). The method proposed was similar to the studies [14] [19] in which the SVM classifier was used. In [14], the color characteristics of the RGB, HSV, and CIE L*a*b spaces were used, reaching an accuracy of 95.28%, and in [19], texture characteristics of the grayscale were used, achieving an accuracy of 95%. The presented method has provided novel information on how color spaces can

influence the selection of features for image analysis of leukemic cells. Future research could extend the classification approach by considering other cancer types or subtypes, using other classifiers, or selecting different feature selection methods.

AUTHOR CONTRIBUTIONS

C.E.-D.A. and A.F.-D. Conceptualized the project, developed aims and goals, participated in software development. C.E.-D.A. Collected data and performed data curation and carried out the statistical analyses, programming, and code implementation; designed and developed the methodology, performed experiments, verified the reproducibility of the results, and oversaw the presentation and visualization of data, contributed to writing the draft and final version of the manuscript, edited critical reviews and responses to reviewer comments. A.F-.D. Provided study materials, literature and computer resources and tools for image processing, provided and work with the leukemia imaging database, oversaw the writing of the manuscript, carried out techniques of software image processing to analyze data, developed and designed the methodology, validated the results of the investigation, obtained funding. Both authors reviewed and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank to the Consejo Nacional de Ciencia y Tecnología (CONACYT) for financing this work through the "Programa Nacional de Posgrados de Calidad (PNPC)" scholarship 1033981.

ETHICAL STATEMENT

The data used in this work is from public dataset: Gupta A, Gupta R. SN-AM Dataset: White Blood Cancer Dataset of B-ALL and MM for Stain Normalization [Internet]. The Cancer Imaging Archive; 2019. Available from: *https://doi.org/10.7937/tcia.2019.of2w8lxr*.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] McKenzie SB, Williams JL. Clinical Laboratory Hematology. 3rd ed. Boston: Pearson; 2014. 1037p.

[2] Bozzone DM. The Biology of Cancer: Leukemia. New York, N.Y.: Chelsea House Pub; 2009. 168p.

[3] Sabath DE. Leukemia. In: Maloy S, Hughes K (eds). Brenner's Encyclopedia of Genetics [Internet]. Academic Press;2013. 226-227p. Available from: https://doi.org/10.1016/B978-0-12-374984-0.00862-7

[4] Halim NHA, Mashor MY, Hassan R. Automatic Blasts Counting for Acute Leukemia Based on Blood Samples. Int J Res Rev Comput Sci [Internet]. 2011;2(4):971-976. Available from: *https://www.lumenera.com/media/wysiwyg/documents/whitepapers/IJRRCS-Research-Article.pdf*

[5] Hazra T, Kumar M, Tripathy SS. Automatic Leukemia Detection Using Image Processing Technique. Int J Latest Technol Eng Manag Appl Sci [Internet]. 2017;6(4):42-45. Available from: *https://www.ijltemas.in/DigitalLibrary/Vol.6Issue4/42-45.pdf*

[6] Putzu L, Caocci G, Di Ruberto C. Leucocyte classification for leukaemia detection using image processing techniques. Artif Intell Med [Internet]. 2014;62(3):179-191. Available from: https://doi.org/10.1016/j.artmed.2014.09.002

[7] Mittal A, Dhalla S, Gupta S, Gupta A. Automated analysis of blood smear images for leukemia detection: a comprehensive review. ACM Comput Surv [Internet]. 2022;1-36. Available from: https://doi.org/10.1145/3514495

[8] Shah A, Naqvi SS, Naveed K, Salem N, et al. Automated Diagnosis of Leukemia: A Comprehensive Review. IEEE Access [Internet]. 2021;9:132097-132124. Available from: https://doi.org/10.1109/ACCESS.2021.3114059

[9] Mohammed ZF, Abdulla AA. Thresholding-based White Blood Cells Segmentation from Microscopic Blood Images. UHD J Sci Technol [Internet]. 2020;4(1):9-17. Available from: https://doi.org/10.21928/uhdjst.v4n1y2020.pp9-17

[10] Alsalem MA, Zaidan AA, Zaidan BB, Hashim M, et al. A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations. Comput Methods Programs Biomed [Internet]. 2018;158:93-112. Available from: https://doi.org/10.1016/j.cmpb.2018.02.005

[11] Anilkumar KK, Manoj VJ, Sagi TM. A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of Leukemia. Biocybern Biomed Eng [Internet]. 2020;40(4):1406-1420. Available from: https://doi.org/10.1016/j.bbe.2020.08.010

[12] Mughal TI, Goldman JM, Mughal ST. Understanding Leukemias, Lymphomas and Myelomas. 2nd ed. London: CRC Press; 2013. 200p.

[13] Dese K, Raj H, Ayana G, Yemane T, et al. Accurate Machine-Learning-Based classification of Leukemia from Blood Smear Images. Clin Lymphoma Myeloma Leuk [Internet]. 2021;21(11):903-914. Available from: https://doi.org/10.1016/j.clml.2021.06.025

[14] Saeedizadeh Z, Mehri Dehnavi A, Talebi A, Rabbani H, et al. Automatic recognition of myeloma cells in microscopic images using bottleneck algorithm, modified watershed and SVM classifier. J Microsc [Internet]. 2016;261(1):46-56. Available from: https://doi.org/10.1111/jmi.12314

[15] P R, P SD. Detection of Blood Cancer-Leukemia using K-means Algorithm. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) [Internet]. Madurai: IEEE; 2021:838-842. Available from: https://doi.org/10.1109/ICICCS51141.2021.9432244

[16] Soni F, Sahu L, Getnet ME, Reta BY. Supervised Method for Acute Lymphoblastic Leukemia Segmentation and Classification Using Image Processing. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) [Internet]. Tirunelveli: IEEE; 2018:1075-1079. Available from: https://doi.org/10.1109/ICOEI.2018.8553937

[17] Jagadev P, Virani HG. Detection of leukemia and its types using image processing and machine learning. In: 2017 International Conference on Trends in Electronics and Informatics (ICEI) [Internet]. Tirunelveli: IEEE; 2017:522-526. Available from: https://doi.org/10.1109/ICOEI.2017.8300983

[18] Kumar P, Udwadia SM. Automatic detection of acute myeloid leukemia from microscopic blood smear image. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) [Internet]. Udupi: IEEE; 2017:1803-1807. Available from: https://doi.org/10.1109/ICACCI.2017.8126106

[19] Mirmohammadi P, Ameri M, Shalbaf A. Recognition of acute lymphoblastic leukemia and lymphocytes cell subtypes in microscopic images using random forest classifier. Phys Eng Sci Med [Internet]. 2021;44(2):433-441. Available from: https://doi.org/10.1007/s13246-021-00993-5

[20] Abdeldaim AM, Sahlol AT, Elhoseny M, Hassanien AE. Computer-Aided Acute Lymphoblastic Leukemia Diagnosis System Based on Image Analysis. In: Hassanien A, Oliva D (eds). Studies in Computational Intelligence [Internet]. Cham: Springer; 2018:730.131-147p. Available from: https://doi.org/10.1007/978-3-319-63754-9_7

[21] Rahman A, Hasan MM. Automatic Detection of White Blood Cells from Microscopic Images for Malignancy Classification of Acute Lymphoblastic Leukemia. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET) [Internet]. Dhaka: IEEE; 2018:1-6. Available from: https://doi.org/10.1109/CIET.2018.8660914

[22] Shafique S, Tehsin S, Anas S, Masud F. Computer-assisted Acute Lymphoblastic Leukemia detection and diagnosis. In: 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE) [Internet]. Islamabad: IEEE; 2019:184-189. Available from: https://doi.org/10.1109/C-CODE.2019.8680972

[23] Singhal V, Singh P. Texture Features for the Detection of Acute Lymphoblastic Leukemia. In: Satapathy S, Joshi A, Modi N, Pathak N (eds). Advances in Intelligent Systems and Computing [Internet]. Singapore: Springer; 2016:535-43. Available from: https://doi.org/10.1007/978-981-10-0135-2_52

[24] Rehman A, Abbas N, Saba T, Rahman SIU, et al. Classification of acute lymphoblastic leukemia using deep learning. Microsc Res Tech [Internet]. 2018;81(11):1310-1317. Available from: https://doi.org/10.1002/jemt.23139

[25] Rawat J, Singh A, HS B, Virmani J, et al. Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia. Biocybern Biomed Eng [Internet]. 2017;37(4):637-654. Available from: https://doi.org/10.1016/j.bbe.2017.07.003

[26] Muntasa A, Yusuf M. Color-Based Hybrid Modeling to Classify the Acute Lymphoblastic Leukemia. Int J Intell Eng Syst [Internet]. 2020;13(4):408-422. Available from: https://doi.org/10.22266/ijies2020.0831.36

[27] Mandal S, Daivajna V, V R. Machine Learning based System for Automatic Detection of Leukemia Cancer Cell. In: 2019 IEEE 16th India Council International Conference (INDICON) [Internet]. New Delhi: IEEE; 2019:1-4. Available from: https://doi.org/10.1109/INDICON47234.2019.9029034

[28] Acharya V, Kumar P. Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms. Med Biol Eng Comput [Internet]. 2019;57(8):1783-1811. Available from: https://doi.org/10.1007/s11517-019-01984-1

[29] Bagasjvara RG, Candradewi I, Hartati S, Harjoko A. Automated detection and classification techniques of Acute leukemia using image processing: A review. In: 2016 2nd International Conference on Science and Technology-Computer (ICST) [Internet]. Yogyakarta: IEEE; 2016:35-43. Available from: https://doi.org/10.1109/ICSTC.2016.7877344

[30] Belhekar A, Gagare K, Bedse R, Bhelkar Y, et al. Leukemia Cancer Detection Using Image Analytics : (Comparative Study). In: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA) [Internet]. Pune: IEEE; 2019:1-6. Available from: https://doi.org/10.1109/ICCUBEA47591.2019.9128546

[31] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference [Internet]. London: IEEE; 2014:372-378. Available from: https://doi.org/10.1109/SAI.2014.6918213

[32] Kumar D, Jain N, Khurana A, Mittal S, et al. Automatic Detection of White Blood Cancer From Bone Marrow Microscopic Images Using Convolutional Neural Networks. IEEE Access [Internet]. 2020;8:142521-142531. Available from: https://doi.org/10.1109/ACCESS.2020.3012292

[33] Sahlol AT, Abdeldaim AM, Hassanien AE. Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm. Soft Comput [Internet]. 2019;23(15):6345-6360. Available from: https://doi.org/10.1007/s00500-018-3288-5

[34] Sahlol AT, Kollmannsberger P, Ewees AA. Efficient Classification of White Blood Cell Leukemia with Improved Swarm Optimization of Deep Features. Sci Rep [Internet]. 2020;10(1):2536. Available from: https://doi.org/10.1038/s41598-020-59215-9

[35] Mishra S, Majhi B, Sa PK. Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection. Biomed Signal Process Control [Internet]. 2019;47:303-311. Available from: https://doi.org/10.1016/j.bspc.2018.08.012

[36] Pešić I. Segmentation and Classification of Leucocyte Images for Detection of Acute Lymphoblastic Leukemia. In: 2020 7th ETRAN&IcETRAN international conference [Internet]. Belgrade: IcETRAN; 2020:2-7. Available from: *https://www.etran.rs/2020/ZBORNIK_RADOVA/Radovi_prikazani_na_konferenciji/047_BTI1.7.pdf*

[37] Mirmohammadi P, Taghavi A, Ameri A. Automatic Recognition of Acute Lymphoblastic Leukemia Cells from Microscopic Images. Int J Innov Res Sci Eng [Internet]. 2017;5(7):8-11. Available from: *https://ijirse.in/docs/2017/Sep%2017/IJIRSE170902.pdf*

[38] Salih Hasan BM, Abdulazeez AM. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. J Soft Comput Data Min [Internet]. 2021;2(1):20-30. Available from: *https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032*

[39] Jolliffe IT. Principal Component Analysis [Internet]. New York: Springer; 2002. 488p. Available from: https://doi.org/10.1007/b98835

[40] Harun NH, Bakar JA, Wahab ZA, Osman MK, et al. Color Image Enhancement of Acute Leukemia Cells in Blood Microscopic Image for Leukemia Detection Sample. In: 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE) [Internet]. Malaysia: IEEE; 2020:24-9. Available from: https://doi.org/10.1109/ISCAIE47305.2020.9108810

[41] Sukanya CM, Vince P. AML Detection in Blood Microscopic Images Using DRLBP and DRLTP Feature Extraction. Int J Eng Sci Comput [Internet]. 2016;6(6):6942-6946. Available from: *https://ijesc.org/upload/16ed93ec7acaf83596e4dc815fc66cad.AML%20Detection%20in%20Blood%20Microscopic%20Images%20Using%20%20DRLBP%20and%20DRLTP%20Feature%20Extraction.pdf*

[42] Rege MV, Abdulkareem MB, Gaikwad S, Gawli BW. Automatic Leukemia Identification System Using Otsu Image segmentation and MSER Approach for Microscopic Smear Image Database. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) [Internet]. Coimbatore: IEEE; 2018:267-272. Available from: https://doi.org/10.1109/ICICCT.2018.8473101

[43] Bhattacharjee R, Saini LM. Detection of Acute Lymphoblastic Leukemia using watershed transformation technique. In: 2015 International Conference on Signal Processing, Computing and Control (ISPCC) [Internet]. Waknaghat: IEEE; 2015:383-386. Available from: https://doi.org/10.1109/ISPCC.2015.7375060

[44] Shinde S, Sharma N, Bansod P, Singh M, et al. Automated Nucleus Segmentation of Leukemia Blast Cells : Color Spaces Study. In: 2nd International Conference on Data, Engineering and Applications (IDEA) [Internet]. Bhopal: IEEE; 2020:1-5. Available from: https://doi.org/10.1109/IDEA49133.2020.9170721

[45] Nor Hazlyna H, Mashor MY, Mokhtar NR, Aimi Salihah AN, et al. Comparison of acute leukemia Image segmentation using HSI and RGB color space. In: 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010) [Internet]. Kuala Lumpur: IEEE; 2010:749-752. Available from: https://doi.org/10.1109/ISSPA.2010.5605410

[46] Inbarani H H, Azar AT, G J. Leukemia Image Segmentation Using a Hybrid Histogram-Based Soft Covering Rough K-Means Clustering Algorithm. Electronics [Internet]. 2020;9(1):188. Available from: https://doi.org/10.3390/electronics9010188

[47] Asadi F, Putra FM, Indah Sakinatunnisa M, Syafria F, et al. Implementation of Backpropagation Neural Network and Blood Cells Imagery Extraction for Acute Leukemia Classification. In: 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME) [Internet]. Bandung: IEEE; 2017:106-110. Available from: https://doi.org/10.1109/ICICI-BME.2017.8537755

[48] Gupta A, Gupta R. SN-AM Dataset: White Blood Cancer Dataset of B-ALL and MM for Stain Normalization [Data set]. The Cancer Imaging Archive; 2019. Available from: https://doi.org/10.7937/tcia.2019.of2w8lxr

[49] Soille P. Morphological Image Analysis: Principles and Applications. 2nd ed. Berlin, Heidelberg: Springer; 2004. 392p.

[50] Moshavash Z, Danyali H, Helfroush MS. An Automatic and Robust Decision Support System for Accurate Acute Leukemia Diagnosis from Blood Microscopic Images. J Digit Imaging [Internet]. 2018;31(5):702-717. Available from: https://doi.org/10.1007/s10278-018-0074-y

[51] Gonzalez RC, Woods RE. Digital Image Processing. 4th ed. New York: Pearson; 2018. 1168p.

[52] R Core Team, R. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna [Internet]. 2016. Available from: *https://www.R-project.org/*

[53] Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Internet]. 2020. Available from: *https://cran.r-project.org/package=factoextra*

[54] Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. 4th ed. New York: Springer; 2002. 516p. Available from: https://doi.org/10.1007/978-0-387-21706-2