How to Cite:
Vidhya, B., & Kombaiya, A. K. (2022). Acute lymphoblastic leukemia prediction using covering-oriented rough K-means segmentation algorithm. *International Journal of Health Sciences*, *6*(S3), 5377–5388. https://doi.org/10.53730/ijhs.v6nS3.7102

Acute lymphoblastic leukemia prediction using covering-oriented rough K-means segmentation algorithm

Mrs. Vidhya B.
Department of Information Technology, Sri Ramakrishna College of Arts & Science, Coimbatore, TamilNadu-641006, India, 9952340679
*Corresponding author email: vidhya.b@srcas.ac.in

Dr. A. Kumar Kombaiya
Department of Computer Science, Chikkana Government Arts College, Tiruppur, TamilNadu-641006, India, 9894838090
Email: kumar_kombaiya@rediffmail.com

***Abstract*---**Acute Lymphoblastic Leukemia (ALL) is a type of dreadful cancer that has an effect on the bone marrow and blood. In recent decades, leukemia mainly affects youngsters and grown-ups. Timely diagnosis of leukemia is crucial to give right treatment to patients, predominantly in the case of children. Computational tecniques for medical image analysis have incredible use and become the focus of research in medical image processing. The major objective of this research work is to predict the acute lymphoblastic leukemia cells using the proposed segmentation algorithms with the assistance of machine learning classifiers. In this study, hybrid clustering based segmentation methods Covering-orientedRough K-Means (CRKM) algorithm is proposed to segment the image of the leukemia nucleus. Then, the performance of the segmentation algorithms were analyzed using different machine learning classifiers. From the experimental analysis, it is observed that the machine learning classification algorithms achieved better prediction accuracy when compared to existing segmentation algorithms.

***Keywords*---**leukemia, segmentation, microscopic image, covering-oriented, rough K-means, nucleus segmentation.

Introduction

Medical imaging is a technique or a process for analyzing the body parts, organs and tissues to diagnose the disease. Image processing and computational

Manuscript submitted: 9 March 2022, Manuscript revised: 18 April 2022, Accepted for publication: 1 May 2022

intelligence techniques produce medical images in digitized form which can be analyzed and studied with the aid of computers. Medical images in digital format facilitate in-depth analysis resulting in a more accurate diagnosis. Leukemia is a kind of malignant tumour which influences the bone marroe and blood. The reason behind the leukemia is the anamalus production of white blood cells (WBCs) in the human body. Blood encompasses millions of cells in the form of liquid which comprises the various components of blood cells namely Red Blood Cells (RBCs), WBCs, plasma and platelets. In human body, WBCs constitute only 1% of blood content which fights against infection. These WBCs are the most affected constituents by cancer. In blood volumes, 40% - 45% of RBCs carry oxygen to the body. Platelets are smaller cells that help to control bleeding. The liquid portion of the blood is called plasma. It helps to transport water and nutrients to the tissues. All the blood components that are made in blood factory are collectively called as the Bone Marrow. In account of leukemia, bone marrow creates a lot of irregular white blood cells that are not completely developed [1]. Figure 1 shows the microscopic images of normal blood and blood infected with leukemia.

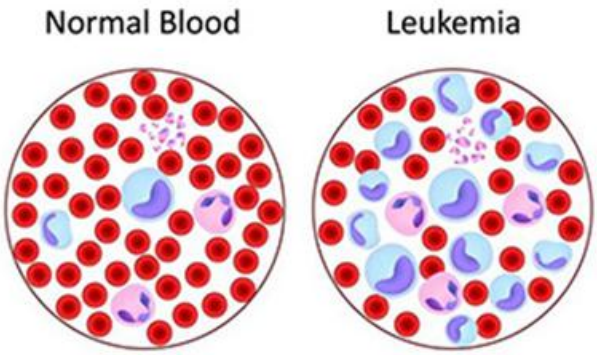

Figure 1. Microscopic Images of Normal Blood and Blood Infected with Leukemia

Classification of leukemia can be made in two ways: Firstly, depending on how fast the cells grow, it can be classified as acute and chronic. In acute leukemia, the abnormal WBCs grow quickly and it is the most common type observed in childhood stage [2]. In chronic leukemia, the blast cells grow slowly. It is mostly seen in older people, but it could also affect persons of any age group. Secondly, depending on the type of WBCs that have turned into cancerous cells, leukemia can be classified as lymphoid or myeloid. Due to the development of superior medical imaging modalities, the manual medical image analysis becomes more difficult. Therefore, the sophisticated and proficient computer aided system is required to study and interpret the images, and diagnose the diseases. Leukemia is a kind of malignant tumour which influences the bone marroe and blood. Acute and chronic are the two broad categories of leukemia [3]. Acute leukemia is classified into two subtypes namely, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL).

ALL is a childhood cancer that primarily affects the children of 5 to 14 years old. If leukemia is predicted and identified in its early stage, an appropriate treatment can be given to the patient; by this way, the disease can be cured, and patient's recovery can be expedited and lives of many leukemia patients can be saved. Leukemia cells are of complex nature. Diagnosing leukemia disease through medical images is a difficult task. The major constraint in leukemia diagnosis is: Non-blast and blast cells are identified according to the shape and size of the image of leukemia nucleus. So nucleus segmentation in the image is significantly important since the accuracy of feature extraction, selection and image classification is entirely subject to the accurate segmentation of the image of the leukemia nucleus. In this research work, a Covering-orientedRough K-Means (CRKM) clustering algorithm is proposed to segment the nucleus of leukemia image. In this method, Covering-orientedapproximation is employed instead of rough approximation. Then the segmented images produced by the existing segmentation algorithms and the proposed CRKM algorithm are deputed to the classification algorithms to verify the significance of the proposed CRKM algorithm.

Related Works

The research study presented in [4] employed segmentation on Prostate MRI images. In preprocessing, stick filtering and contrast stretching were applied to remove the noise in an image. In this research, Fuzzy C-Means (FCM) clustering algorithm was utilized for segmenting the image of the prostate gland from the axial MRI slices. This approach automatically segmented the ROI of the prostate image. In the research work presented in [5] introduced an approach involving Soft Fuzzy Rough C-Means (SFRCM) clustering algorithm for Magnetic Resonance Imaging (MRI) brain tumour image segmentation. This clustering algorithm was formulated by incorporating the concepts of soft set, fuzzy set, and rough set theory. The fuzzified values were used to compute the approximations of the cluster instead of crisp values. In this approach the centroids were selected by using histogram of the image.

In [6] the authors presented Fuzzy Spectral Clustering (FSC) algorithm to segment images of wound regions. This approach was adopted to demarcate the ulcer boundary in the image. The proposed method was validated using the ground-truth images labelled by the two skilled dermatologists. The FSC algorithm produced better results in image segmentation. Su et al. [11] developed a diagnostic system for acute myeloid leukemia (AML) images. This system was purely based on the counting of blast cells in the blood samples. The nucleus and cytoplasm in the images were extracted using Hidden-Markov Random Field (HMRF). Table 1 provides a summary of the characteristics of different clustering algorithms applied for image segmentation.

Table 1
Summary of Clustering-Based Segmentation Algorithms

Authors	Year	Imaging Modality	Type of Disease	Preprocessing Techniques	Clustering Algorithms
Rundo et al. [7]	2018	MRI	Prostate cancer	Stick Filtering	Fuzzy C-Means (FCM)

Namburu et al. [8]	2017	MRI	Brain tumor	Medical Image Processing, Analysis and Visualization (MIPAV)	Soft Fuzzy Rough C-Means algorithm
Tosta et al. [9]	2017	Microscopic image	Leukemia	Histogram Equalization, Gaussian Filter	Fuzzy 3-partition entropy and genetic algorithm
Dhane et al. [10]	2017	Digital camera	Chronic wound	Ordered Statistics Filter	Fuzzy spectral clustering, Gray-based fuzzy similarity measure
Su et al. [11]	2017	Microscopic image	Acute myeloid leukemia	Lab Colour Space	K-Means, Hidden-Markov RandomField
Kaya et al. [12]	2017	MRI	Brain tumor	Box Filtering	Principle Component Analysis (PCA), K-Means, Fuzzy C-Means

Methodology

The steps followed in acute lymphoblastic leukemia image classification are showed in Figure 2. This approach comprises of image acquisition, preprocessing, nucleus segmentation, and classification.

Image Acquisition and Preprocessing

The acute lymphoblastic leukemia image databases namely, ALL-IDB1 and ALL-IDB2 are the datasets of digital microscopic images of blood samples. It is extensively used for evaluating and comparing of segmentation and classification algorithms. The considered data is made available from the webpage https://homes.di.unimi.it/scotti/all/ (Labati et al. 2011; Scotti, 2005). Medical images are obtained from different modalities such as MRI, X-Ray, CT scan, and microscope, which is very hard to handle directly. Therefore, preprocessing techniques are used to remove the noise and enhance the quality of an image. In this research work, the pre-processing techniques include noise removal and contrast enhancement. The digital microscopic images are often corrupted by noise and speckles due to the fault of electronic equipment, environmental factor, and human error during the process. Inference from noise-affected images may lead to imperfect diagnosis. In such instances, medical image processing helps a lot in producing clear images. In the current research, popular filtering techniques are employed like median filter, average filter, and Gaussian filter to enhance the image of the leukemia cell.

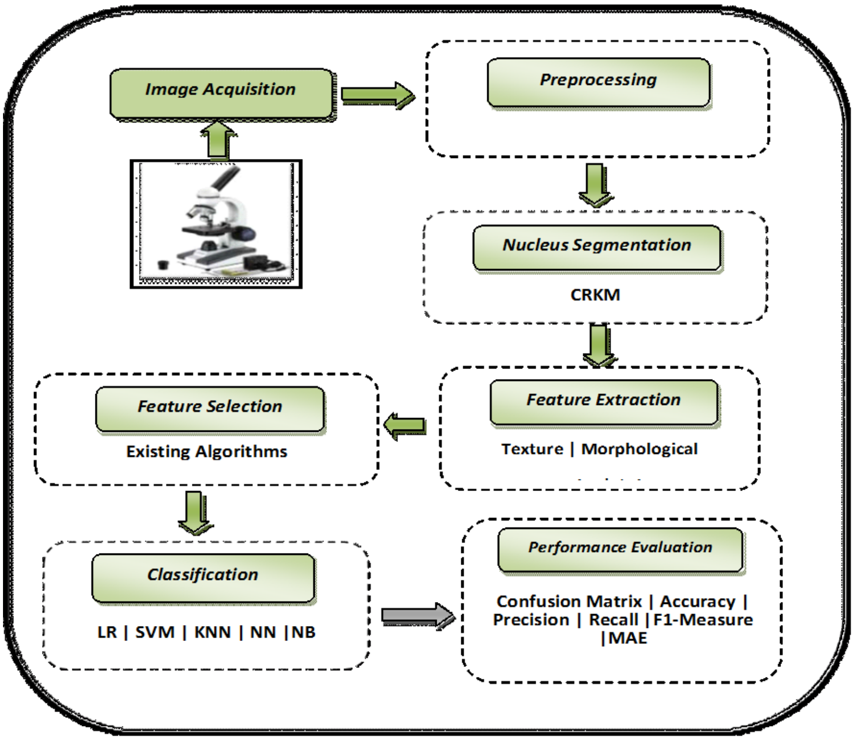

Figure 2. Methodology of ALL Image Classification

Average filter is a sliding-window spatial filter which replaces each pixel value of an image with the average value of the entire pixel values in the window. Median filtering is an efficient filtering technique, widely applied for removing 'salt and pepper' kind of noise in the images. In median filtering, every pixel of an image is substituted with the median value of neighbouring pixels. Gaussian filter is used to remove the noise of an image using Gaussian function. Thus, the images of leukemia cells were preprocessed and fed as input for further processing, into image segmentation process.

Proposed Covering-Oriented Rough K-Means Clustering

Covering-oriented rough set is an expansion of existing theory which alters rough approximation, partition or similarity relation. Covering-oriented rough set differs from the classical rough set, which uses covering of the universal set instead of partitions or equivalence class. The covering created by symmetric and reflexive relations and it finds the similarity of the unary covering. The property of covering is the intersection of any two features is the blending of finite features. This algorithm incorporates the Covering-orientedrough approximation for computing the upper and lower estimations of the clustering algorithm. The CRKM clustering algorithm is outlined in Algorithm.

Algorithm : Covering-Oriented Rough K-Means Clustering

Input: Img, K, w_l, w_b, u.
Output: *Segmented Image* ($SegImg$)

***Step* 1**. Initialize the parameters K, w_l, w_b, u.
***Step* 2**. Assign each pixel randomly to exactly one lower approximation of the clusters.
***Step* 3**. Calculate the new means m_k for every cluster C_k.
***Step* 4**. Assign the pixel to the approaximation and identify the closest mean m_h.
$d_{n,h}^{min} = d(X_n, m_h) = \min d(X_n, m_k)$
***Step* 5**. If $\frac{d(X_n,m_h)}{d(X_n,m_k)} \le v$ *where* $h \ne k$ *Then* $Xn \in \overline{C}_t$
Else $Xn \in \underline{C}_h$
***Step* 6**. Repeat Steps 2-4 until the absence of new object assignments
***Step* 7**. Return *SegImg*

First, initialize all the parameters viz., number of clusters k, lower weight w_l, upper weight w_b, and the threshold u. Each pixel is assigned into one lower estimation of the cluster. The new mean m_k for each cluster C_k is computed using:

$$m_k = \begin{cases} w_l \sum_{X_n \in \underline{C_k}_*} \frac{X_n}{|\underline{C_k}_*|} + w_b \sum_{X_n \in C_k^b} \frac{X_n}{|C_k^b|} & for\ C_k^b \ne \phi \\ w_l \sum_{X_n \in \underline{C_k}_*} \frac{X_n}{|\underline{C_k}_*|} & otherwise, \end{cases}$$

Data pertaining to each pixel are assigned to their closest mean. Based on the given threshold, each pixel is assigned to the corresponding upper or lower approximation of the cluster. This algorithm is applied until all the data objects close to cluster remain unchanged. Finally, the clustered image is labelled by cluster index and the segmented imageof the nucleus is extracted.

Feature Extraction, Feature Selection and Classification

In this research, three different types of features were employed to address the segmented nucleus image, which includes shape, colour and texture-based features. In texture features, Grey Level Co-occurrence Matrix (GLCM) was calculated for various dimensions. A total of 108 features consisting of 8 shape-based features, 87 GLCM texture features (each direction 23 features) and 12-colour based features were extracted. The Supervised Quick Reduct (SQR) algorithm selects 73 features (67%) as prominent features. Then the selected features are imputed to the classification algorithms. In this experiment, five machine learning algorithms namely, Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbourhood (KNN), Support Vector Machine (SVM), and Neural etwork (NN) were employed, to assess the performance of the proposed segmentation CRKM algorithm [13].

Experimental Results and Discussion

The output results of the images Im148_0, Im212_0, Im100_1, Im072_1,and Im154_0 after the application of the proposed CRKM clustering based segmentation algorithms are displayed in Figure 3. The original image, clustered image, cluster1, cluster2, and cluster3 images are depicted in Figure 3. The performance of the machine learning classifiers were analyzed using various metrics like Accuracy (A), Recall (R), Precision (P), Mean Absolute Error (MAE), and F1 Measure. [14]. The classification results with respect to the K-means algorithm, RKM segmentation algorithm and proposed CRKM segmentation algorithm are presented in Tables 2 and 3 respectively. The performance of the segmentation algorithms were analyzed using different machine classifiers. The experimental results illustrates that the proposed CRKM algorithm performs better than the other algorithms, whose results are represented graphically in Figure 4, 5, and 6 respectively.

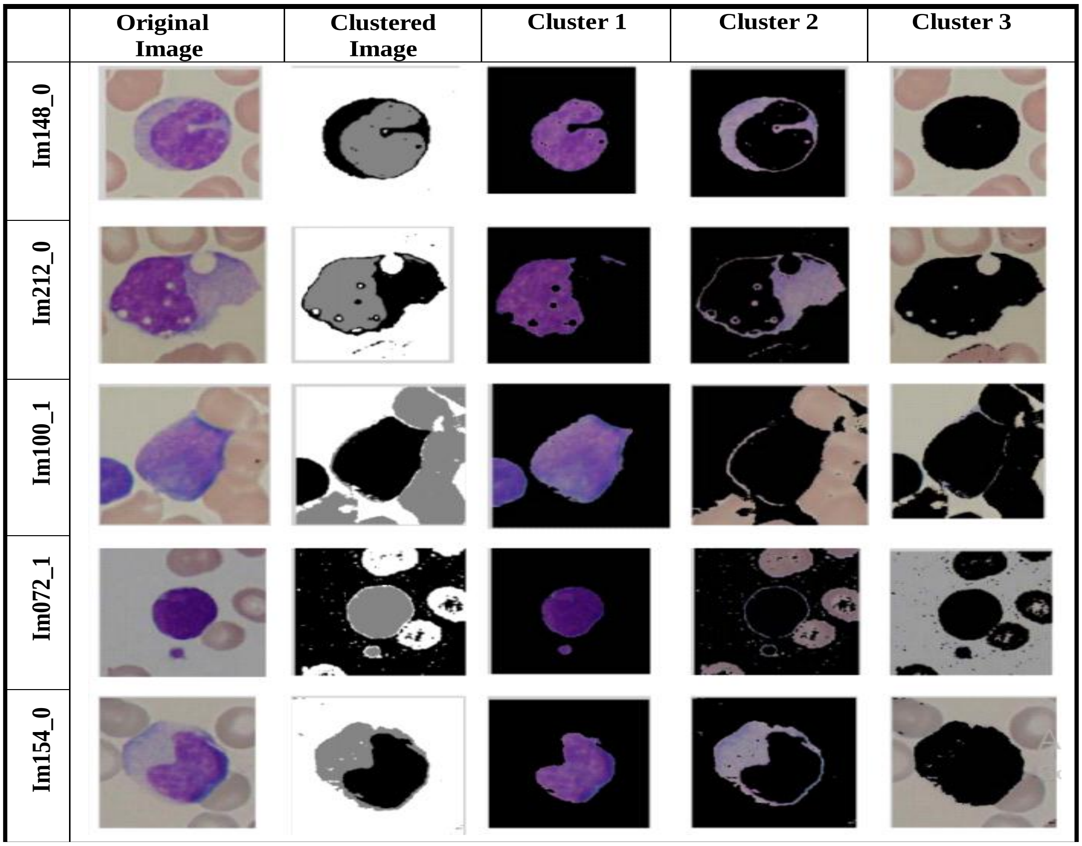

Figure 3. Segmentation Results of Proposed CRKM

Table 2
Performance Analysis of K-Means Clustering

ML Algorithms	Dataset	P	R	F1	MAE	A
	GLCM-0	81.00	77.00	78.00	0.134	76.74
	GLCM045	79.00	79.00	78.00	0.087	79.07
LR	GLCM-90	60.00	66.00	68.00	0.112	65.17
	GLCM-135	70.00	66.00	67.00	0.128	67.44

	SC	76.00	74.00	75.00	0.115	74.42
	GLCM-0	61.00	60.00	61.00	0.082	60.47
	GLCM045	62.00	61.00	58.00	0.093	60.60
NB	GLCM-90	54.00	51.00	50.00	0.128	51.16
	GLCM-135	60.00	56.00	57.00	0.088	55.81
	SC	68.00	65.00	65.00	0.110	65.11
	GLCM-0	77.00	74.00	71.00	0.073	74.41
	GLCM045	80.00	72.00	72.00	0.113	72.09
SVM	GLCM-90	82.00	70.00	73.00	0.032	69.77
	GLCM-135	65.00	65.00	65.00	0.130	65.11
	SC	73.00	70.00	67.00	0.113	69.78
	GLCM-0	71.00	72.00	72.00	0.119	72.09
	GLCM045	71.00	67.00	39.00	0.062	67.44
KNN	GLCM-90	69.00	67.00	67.00	0.138	67.44
	GLCM-135	63.00	63.00	63.00	0.135	62.79
	SC	67.00	67.00	67.00	0.112	67.44
	GLCM-0	79.00	79.00	76.00	0.077	79.06
	GLCM045	77.00	77.00	77.00	0.139	76.74
NN	GLCM-90	66.00	67.00	69.00	0.094	66.66
	GLCM-135	72.00	70.00	71.00	0.105	69.69
	SC	65.00	65.00	64.00	0.116	65.11

Table 3
Performance Analysis of RKM-based Clustering

ML Algorithms	Dataset	P	R	F1	MAE	A
LR	GLCM-0	82.00	81.00	82.00	0.141	81.39
	GLCM045	88.00	86.00	85.00	0.150	86.04
	GLCM-90	80.00	79.00	76.00	0.143	79.06
	GLCM-135	90.00	86.00	86.00	0.065	86.04
	SC	81.00	81.00	82.00	0.092	81.39
NB	GLCM-0	69.00	67.00	68.00	0.082	67.44
	GLCM045	60.00	64.00	66.00	0.118	63.63
	GLCM-90	56.00	61.00	58.00	0.093	60.61
	GLCM-135	64.00	64.00	65.00	0.012	63.63
	SC	61.00	62.00	62.00	0.118	62.69
SVM	GLCM-0	80.00	79.00	76.00	0.143	79.07
	GLCM045	79.00	79.00	78.00	0.085	79.06
	GLCM-90	71.00	72.00	71.00	0.105	72.09
	GLCM-135	76.00	77.00	72.00	0.086	76.74
	SC	81.00	81.00	79.00	0.120	81.39
KNN	GLCM-0	82.00	79.00	80.00	0.123	79.06
	GLCM045	71.00	73.00	71.00	0.082	72.72
	GLCM-90	70.00	67.00	68.00	0.141	67.44
	GLCM-135	75.00	74.00	73.00	0.176	74.41
	SC	82.00	81.00	81.00	0.118	80.66
	GLCM-0	62.00	76.00	68.00	0.075	75.44

NN	GLCM045	61.00	73.00	66.00	0.123	72.12
	GLCM-90	88.00	84.00	83.00	0.053	83.72
	GLCM-135	91.00	91.00	91.00	0.062	90.67
	SC	86.00	86.00	86.00	0.070	86.12

Table 4
Performance Analysis of CRKM Segmentation Algorithm

ML Algorithms	Dataset	P	R	F1	MAE	A
LR	GLCM-0	84.00	84.00	85.00	0.017	84.72
	GLCM045	93.00	93.00	93.00	0.072	93.02
	GLCM-90	87.00	86.00	86.00	0.032	87.65
	GLCM-135	89.00	88.00	88.00	0.112	88.37
	SC	86.00	85.00	85.00	0.047	85.65
NB	GLCM-0	70.00	70.00	70.00	0.190	69.76
	GLCM045	67.00	65.00	65.00	0.128	65.11
	GLCM-90	67.00	65.00	65.00	0.128	65.11
	GLCM-135	61.00	58.00	56.00	0.152	58.13
	SC	84.00	84.00	85.00	0.017	84.72
SVM	GLCM-0	84.00	81.00	81.00	0.140	81.39
	GLCM045	84.00	81.00	81.00	0.140	81.36
	GLCM-90	79.00	79.00	79.00	0.321	79.06
	GLCM-135	74.00	74.00	73.00	0.132	74.41
	SC	86.00	84.00	84.00	0.089	83.92
KNN	GLCM-0	86.00	84.00	84.00	0.072	83.92
	GLCM045	82.00	81.00	81.00	0.097	81.39
	GLCM-90	75.00	72.00	71.00	0.127	72.09
	GLCM-135	77.00	77.00	76.00	0.101	76.74
	SC	86.00	85.00	85.00	0.047	85.65
NN	GLCM-0	86.00	86.00	86.00	0.070	86.04
	GLCM045	91.00	91.00	90.00	0.054	90.69
	GLCM-90	84.00	84.00	85.00	0.017	84.72
	GLCM-135	93.00	93.00	93.00	0.074	93.02
	SC	86.00	87.00	86.00	0.047	85.65

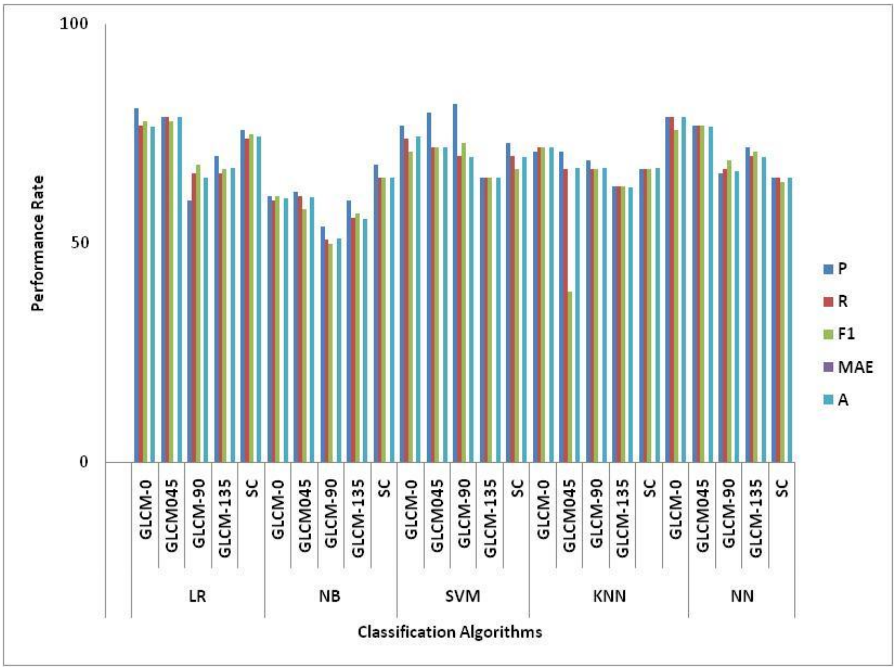

Figure 4. Performance Analysis of K-Means Clustering Algorithm

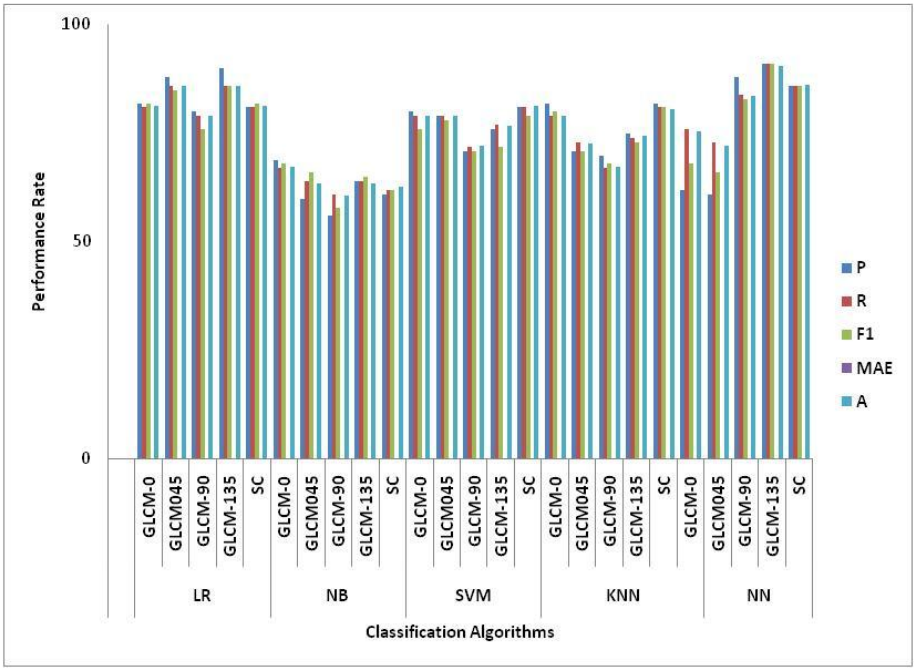

Figure 5. Performance Analysis of RKM-Based Clustering Algorithm

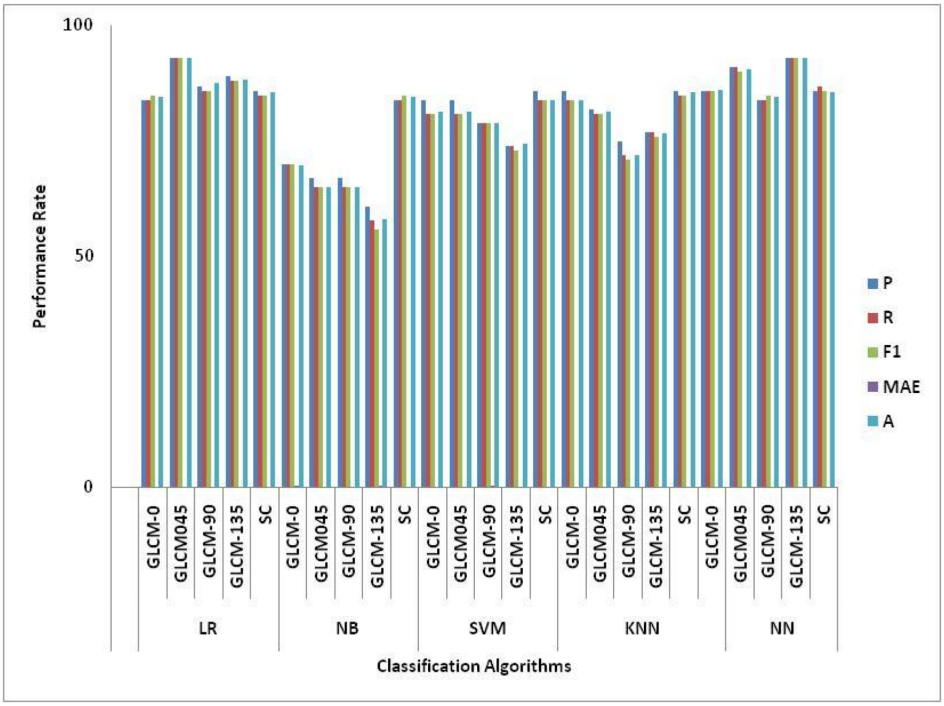

Figure 6. Performance Analysis of CRKM Segmentation Algorithm

Conclusion

Leukemia is a type of cancer affecting blood-forming tissues, blood and bone marrow. Despite advancements in medical imaging modalities it remains difficult to analyze the medical images accurately by physical examination alone. Therefore, a proficient computer aided system is required to study the images and diagnose the diseases. The main objective of this research is to predict acute lymphoblastic leukemia at its early stage from images by applying computational intelligence techniques involving appropriate and efficient algorithms. In the proposed CRKM algorithm, Covering-oriented rough set approximation has been used to find the lower and upper approximation of the clusters. The performance of the segmentation algorithms were analyzed using machine learning classifiers. With respect to CRKM algorithm, most of the ML methods (except Naive Bayes) achieved greater than 80% prediction accuracy as compared when compared to existing segmentation algorithms like K-Means clustering and Rough K-Means clustering. In future, a Computer Aided Design (CAD) system could be developed to study the medical images more clearly and to predict the leukemia using deep learning concept.

References

1. Bracher-Smith M, Crawford K, Escott-Price V (2020) Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol Psychiatry:1–10.

2. Taub FE, DeLeo JM, Thompson EB (1983) Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. DNA 2(4):309–327.
3. Goswami RS, Sukhai MA, Thomas M, Reis PP, Kamel-Reid S (2009) Applications of microarray technology to acute myelogenous leukemia. Cancer informatics 7:CIN. S1015.
4. Ms. A.Sakila and Dr. S.Vijayarani, "Image Enhancement using Morphological Operations", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN: 2395-1990, Online ISSN: 2394-4099, Volume 3, Issue 2, April 2017.
5. Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 23(1):89–109.
6. Mahmood N, Shahid S, Bakhshi T, Riaz S, Ghufran H, Yaqoob M (2020) Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. Medical & Biological Engineering & Computing:1–10.
7. Rundo, L., Militello, C., Russo, G., D'Urso, D., Valastro, L. M., Garufi, A Gilardi, M. C. (2018). Fully Automatic Multispectral MR Image Segmentation of Prostate Gland Based on the Fuzzy C-Means Clustering Algorithm. In Multidisciplinary Approaches to Neural Computing, Springer, Cham. 23-37.
8. Namburu, A., kumar Samay, S., Edara, S. R. (2017). Soft fuzzy rough setbased MR brain image segmentation. Applied Soft Computing, 54, 456-466.
9. Tosta, Thaína A. Azevedo, Paulo Rogério Faria, Leandro Alves Neves, and Marcelo Zanchetta do Nascimento. (2017). Computational method for unsupervised segmentation of lymphoma histological images based on fuzzy 3- partition entropy and genetic algorithm. Expert Systems with Applications. 81, 223-243.
10. Dhane, D. M., Maity, M., Mungle, T., Bar, C., Achar, A., Kolekar, M., & Chakraborty, C. (2017). Fuzzy spectral clustering for automated delineation of chronic wound region using digital images. Computers in biology and medicine, 89, 551-56.
11. [Su, Jie, Shuai Liu, and Jinming Song. (2017). A segmentation method based on HMRF for the aided diagnosis of acute myeloid leukemia. Computer Methods and Programs in Biomedicine 152, 115-123.
12. Kaya, I. E., Pehlivanli, A. Ç., Sekizkardes, E. G., Ibrikci, T. (2017). PCA based clustering for brain tumor segmentation of T1w MRI images. Computer Methods and Programs in Biomedicine, 140, 19-28.
13. P. B. Pankajavalli, G. S. Karthick and R. Sakthivel, "An Efficient Machine Learning Framework for Stress Prediction via Sensor Integrated Keyboard Data," in IEEE Access, vol. 9, pp. 95023-95035, 2021.
14. Pankajavalli P.B., Karthick G.S. (2020) A Unified Framework for Stress Forecasting Using Machine Learning Algorithms. In: Chillarige R., Distefano S., Rawat S. (eds) Advances in Computational Intelligence and Informatics. ICACII 2019. Lecture Notes in Networks and Systems, vol 119. Springer, Singapore.