Automated Detection of Acute Lymphocytic Leukemia Using Blast Cell Morphological Features

S. Hariprasath
Assistant Professor, Dept. of ECE
Saranathan College of Engineering
Trichy, India
hariprasath-ece@saranathan.ac.in

T. Dharani
PG Scholar, Dept. of ECE
Saranathan College of Engineering
Trichy, India
dharanithirugnanam@gmail.com

Shaikh Mohammad Bilal N
Assistant Professor, Dept. of CS
KJ Somaiya College
Mumbai, India
mohammadbilal@somaiya.edu

Abstract— **Diagnosing leukemia accurately is a vital issue in the medical field so that it is effectively treated. Abnormalities in the lymphatic system causes 'Leukemia'. Compared to adults, children are more affected by this disease. White Blood cells which are known as potent infection fighters are involved in causing leukemia. Leukemia can affect both lymphocytes and myelocytes in the bone marrow. The major types of leukemia are lymphocytic leukemia and myelocytic leukemia. If the malignant cells are capable of performing their normal biological functions, it is known as Chronic Leukemia or else it is known as Acute Leukemia. In this work, detection of Acute Lymphocytic Leukemia (ALL) is focused as it is more decisive when compared to other types of leukemia. Based on blast cell morphological features, the given samples are classified into benign and malignant classes. The performance of SVM-L, SVM-R and KNN classifiers are analyzed in detection and classification of cells. For the application of noisy training samples, the measured accuracy, sensitivity and specificity are 90.33%, 90%, and 90.9% respectively for KNN classifier whereas for SVM-R classifier the accuracy, sensitivity and specificity are obtained as 91.5%, 90%, and 92% respectively. Based on the results obtained through the experiments and analysis of the results, it is concluded that KNN classifier produces better results for noise free data.**

Keywords— Leukemia, ALL, Pattern Classification system, SVM, KNN.

I. INTRODUCTION

Generally White blood cells has major part in the immune system of our body. Since Leukemia alters the morphology of these cells, WBCs also helps in detecting leukemia. Leukemia/ Blood Cancer is considered as one of the fatal diseases in our country since death rate due to this disease is increasing every year [1].

The French-American-British (FAB) classifies leukemia into Acute and Chronic Leukemia. And also it classifies ALL into ALL-L1, ALL-L2 and ALL-L3. The identification of these disease is necessary in order to provide the appropriate treatment to the patients since ALL may cause immediate death. These subtypes can be identified by their cell nuclear structure and cytoplasm structure. The other types of leukemia are known as Acute Myeloid Leukemia (AML) and Chronic Myeloid Leukemia (CML) that occurs due to abnormal myelocytes [2].

ALL is the common cancer disease which affects mostly the children below age 15. It is a fast growing disease. Hence accurate detection of the disease is necessary for providing the effective treatment to the affected person. By using the morphological features of the blast cell, malignant cells can be distinguished from the benign cells where immature WBCs are known as 'blast cells'. Benign cells and malignant cells can be discriminated by their nuclear structure, nucleus-cytoplasmic ratio, and irregular cell boundaries. The difference between healthy and cancerous cell's structure is shown in figure 1. As seen in the figure 1, cancerous cells have irregular structure and the shape of nucleus is collapsed.

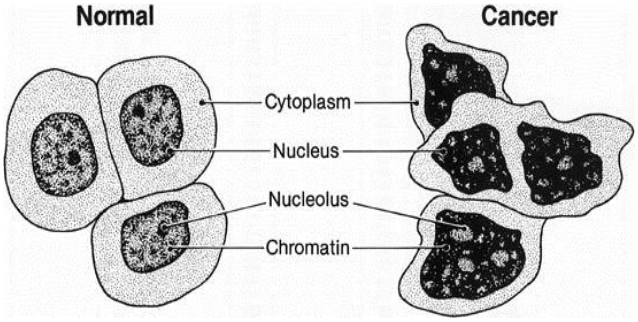

Figure 1. Difference between healthy and cancerous cell structure

II. EXISTING SYSTEMS

Many researchers in medical field have been working in order to obtain the accurate detection of diseases. However the accuracy depends on various factors such as noise free data, features extracted and classifiers used.

Lorentzo Putzu et al has proposed an automatic system to detect and classify leucocytes. The dataset used for this analysis comprises of 108 images from ALL-IDB which is a public image dataset of peripheral blood samples. In this method leucocyte identification is done by using 131 features that consists of shape, color and texture descriptors of a nucleus. Various kernel functions of SVM such as linear, quadratic, polynomial, radial basis functions were analyzed and compared with K-NN, Naive Bayes (NB) classifier and decision tree classifiers. SVM-R achieved accuracy of about 90% (different lightning conditions) when new feature set is used for classification. The sensitivity and specificity are needed to be validated for different lightning conditions [3].

Endah Purwanti and Evelyn Calista has used combination of shape and histogram features for the detection of ALL [4]. He used KNN for classifying the given images (ALL-IDB dataset) into normal and abnormal leucocytes. Different k values are analyzed for the classification and better accuracy of 90% is achieved when the value of k is 7 [4]. And the best combination of feature to achieve this accuracy is area-perimeter-mean-standard deviation.

Bennet Rajesh and S. Sathiamoorthy has integrated genetic algorithm and KNN for noise removal and pre-processing of the given datasets. To reduce the salt and pepper noise, median filter is used. Euclidean distance is used to find the nearest neighbor by KNN for this method [5]. The algorithm is proposed to detect the best value of k with minimum misclassification rate by using genetic algorithm.

Jyoti Rawat Annapurna Singh, H.S. Bhadauria, Jitendra Virmani, Jagtar Singh Devgun, has designed a system that classifies the given samples into healthy and cancerous cells by using a genetic algorithm based SVM. This system uses 331 morphological features and it is tested on 240 microscopic blood smear images (images from American Society of Hematology) for various SVM kernel and achieved an accuracy of about 90%. But this work failed to address the issues of images with irregular illumination [6].

III. METHODLOGY

Steps involved in the proposed methodology is represented in figure 3. Each step is explained in the following sections.

Image Acquisition:

Initial step in the process is to collect the cancer dataset. Generally ALL-IDB dataset consists of 100 images and it is used for cancer detection. It has two distinct versions such as (ALL-IDB1 & ALL-IDB2) [7]. These images are captured by using optical magnifier laboratory as well as Canon Power Shot G5 Camera. The sample images of leucocytes are given in figure 2.

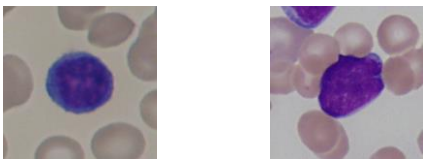

Figure 2. Samples from ALL-IDB for healthy and malignant cells

Image Preprocessing:

Since the given input is in RGB format, the identification of cells is a difficult one. Identification of WBC can be made by converting the RGB format into CMYK format since leucocytes are more concentrated in Y component.

Image Segmentation:

In order to enhance the contrast of the image, histogram equalization or contrast stretching is used. It is done by adjusting the image intensity. Thresholding is fixed automatically to achieve segmentation. It is made by using Zack Thresholding or triangle method.

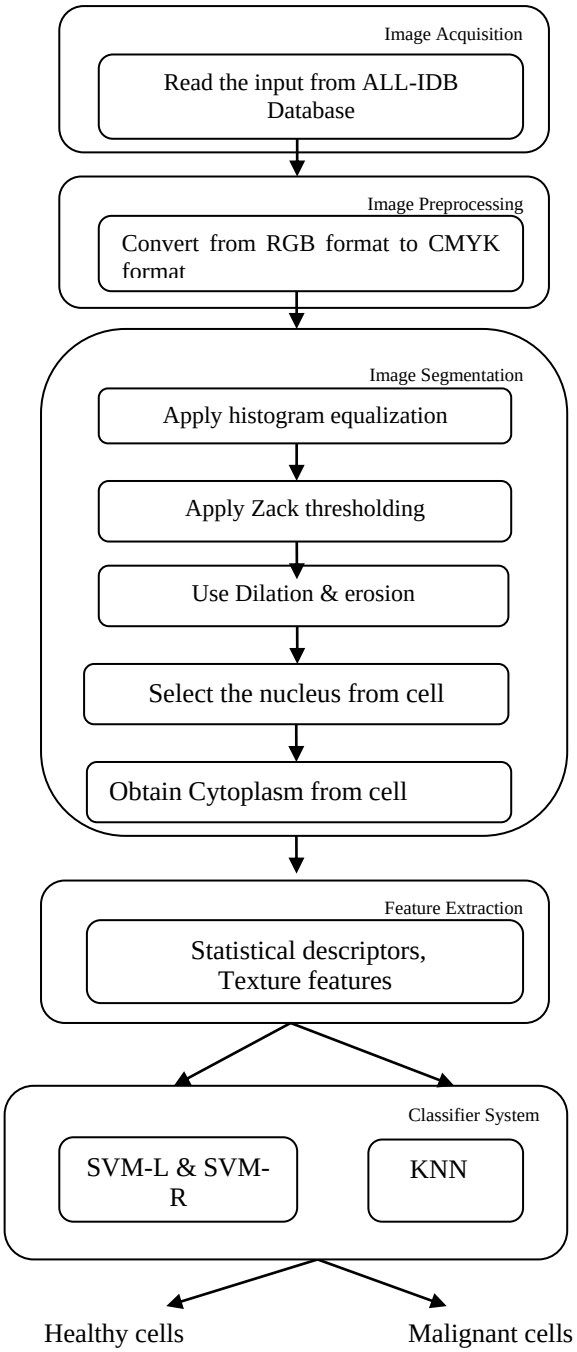

Figure 3. Proposed Methodology

It is applied to the histogram of the image, resulting in a straight line that connects the highest histogram value (h[b max]) and the lowest histogram value (h[b min]), where b max and b min indicate the values of the grey levels where the

histogram h[x] reaches its maximum and minimum, respectively [3]. After applying Zack algorithm to the histogram of the image, threshold is fixed as 0.5664.

Feature Extraction:

Before extracting the features, nucleus and cytoplasm are selected from the cell. The binary image obtained from the green component of RGB format is combined with the binary image obtained from the a* component of the CIE Lab color space via threshold operation. Then, subtraction operation is done between the binary image containing the whole leucocyte and the image containing only the nucleus to obtain the cytoplasm.

Finally, Features such as statistical and texture features are extracted in order to discriminate the benign and malignant cells. Shape [8] and Color descriptors are known as Statistical Descriptors. Nine shape descriptors such as **Area, Eccentricity, Elongation, Solidity, Circularity, Perimeter, Rectangularity, Roundness** and **Orientation.** Color descriptors such as **mean, standard deviation, skew ness, kurtosis and entropy** are calculated from sub images in shades of grey [9] [10]. Twenty Texture features such as auto correlation, contrast, correlation, entropy, energy, homogeneity etc. are extracted from the cell's nucleus.

Classifier System:

By using a classifier, the given set of samples can be classified into healthy and malignant classes. In this work SVM classifier and KNN classifier are used [11].

Support Vector Machine Classifier:

It is a supervised classifier and an optimal hyperplane is used to categorize the given inputs. In 2D plane, a hyperplane is treated as a line that divides a plane into two parts. Mathematically hyperplane can be expressed as,

$$f(x) = \beta_0 + \beta(x)$$

Where β is a weight vector, β0 is a bias.

During implementation, bias is taken as 0.3344. A set of functions that are used in this classifier are called kernel. It provides the required form of the given input. There are different types of kernels such as linear, nonlinear, Radial Basis Functions (RBF), Sigmoid. In this work, SVM-R (RBF) and SVM-L (Linear) is taken for analysis. The mathematical functions of these kernels are given in the table 1, in which x1 and x2 represents the best features selected by SVM.

To validate the SVM, K-Fold Cross Validation technique is used.

K-fold cross validation:

It is a resampling procedure which is used to estimate the skill of machine learning models for new data [12].

Where k is the number of groups that a given sample is to be split into. Generally k is chosen as 5 or 10.

Table 1. Kernel functions of SVM

Kernel function	Formula
Linear	$G(x1,x2) = x1'x2$
RBF/ Gaussian	$G(x1,x2) = \exp(-\|\|x1 - x2\|\|)^2$

K-NN Classifier:

The K-NN is a supervised machine learning classifier. It is based on majority voting process. Choosing the value of K in KNN is a major problem. The value of K depends on the number of training samples. K is normally taken as an odd number. To find nearest neighbor of a given sample can be obtained by calculating the distance. There are various distance metrics such as Euclidean distance, Mahalanobis distance, Minkowski distance etc. for determining nearest neighbor points in KNN.

IV. RESULTS & ANALYSIS

The experiment is done using MATLAB (2017a) which makes easier to compute complex mathematical functions and it is easier to visualize the results [13].

Table 2 shows the texture feature values obtained for various benign and malignant cells. Since healthy and cancerous cells are differed by their morphology, the texture values of both cells are contrasted.

Table 2. Texture features for benign and malignant samples

Type of cell	Auto Correlation	Contrast	Energy	Entropy
Benign	20.24	0.0298	0.4417	1.196
Benign	20.02	0.0302	0.4457	1.23
Benign	19.82	0.0292	0.411	1.35
Malignant	24.12	0.0483	0.2595	1.623
Malignant	24.77	0.0513	0.2635	1.655

Malignant	25.97	0.0657	0.3535	1.756

Classification results:

SVM classifier classifies the given images into two classes namely: healthy and malignant classes. SVM classifier selects best parameters out of selected features. The model is tested with the test set and hyperplane is visualized [14].

Sample points that are very close to the hyperplane is known as 'support vectors'. The results obtained by SVM are shown in figure 4 & 5, in which x axis and y axis represents solidity and standard deviation respectively since SVM selects solidity and standard deviation as best parameters for setting an optimal hyperplane to divide the test samples into two classes.

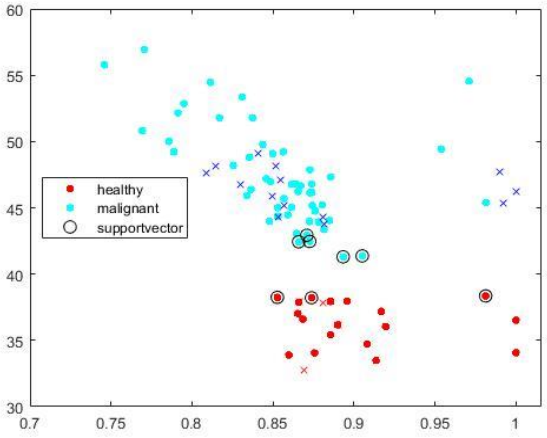

Figure 4. Results obtained- SVM-L

(X-Axis: Solidity, Y-Axis: Standard Deviation)

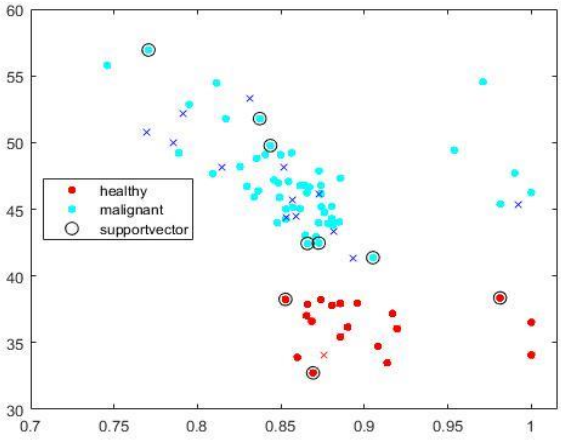

Figure 5. Results obtained- SVM-R

(X-Axis: Solidity, Y-Axis: Standard Deviation)

Implementation of SVM & KNN:

Figure 6 & 7 displays the MATLAB output of SVM-R & SVM-L classifier respectively which shows the hyperplane and also a table which contains true and predicted labels.

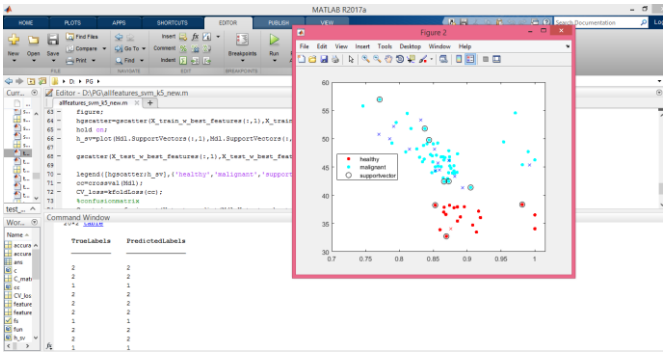

Figure 6. Output of SVM-R showing the hyperplane and predicted labels

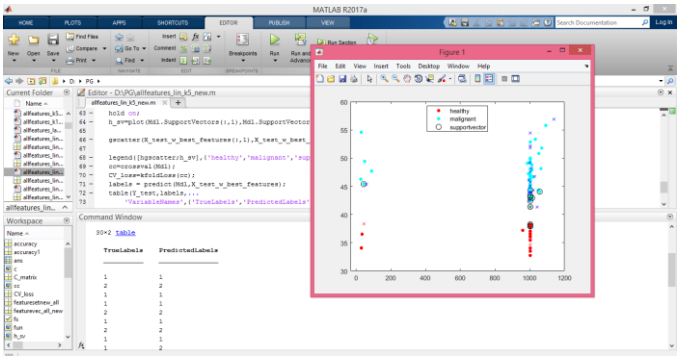

Figure 7. Output of SVM-L showing the hyperplane and predicted labels

Figure 8 displays the KNN output for two cell classification with a table showing the true and predicted labels.

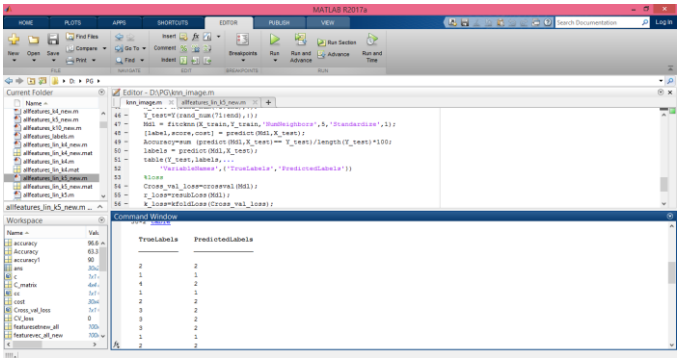

Figure 8. Output of KNN showing true and predicted labels (K=3)

In KNN, 3 nearest neighbors and k value is taken as 3 in our consideration. From the results it can be seen that KNN is well suitable for the classification of benign and malignant cells, only when the dataset used is noise free and small [15].

To calculate the performance metrics such as accuracy, sensitivity and specificity, TP, FP, TN and FN are needed to be found [16].

True positive (TP) = the number of cases correctly identified as malignant.
False positive (FP) = the number of cases incorrectly identified as malignant.
True negative (TN) = the number of cases correctly identified as benign.
False negative (FN) = the number of cases incorrectly identified as benign.

From TP, FP, TN and FN values, performance metrics are obtained using the formula listed in the table 3.

Table 3. Performance metrics and formula

Performance Metrics	Formula
Accuracy	*TP+TN/ (TP+TN+FP+FN).*
Sensitivity	*TP/ (TP+FN).*
Specificity	*TN/ (TN+FP).*

Table 4 shows the results obtained for noise free data. From the results, it can be seen that KNN produces better accuracy compared SVM classifiers. However KNN suffers from high K-fold Loss which is due to the value of K chosen [17].

Table 4: Performance metrics for various classifiers (noise free data) K=5.

Classifier	Sensitivity	Specificity	Accuracy	K-Fold Loss
SVM-L	10%	55%	65%	0.0286
SVM-R	85%	80%	86.67%	0.0125
KNN	90%	90.9%	91.33%	0.0286

When noisy data is analyzed, SVM-R obtained better accuracy (shown in table 5) than KNN classifier. However SVM-L produced poor results in both the cases. The results shown in the table are average of three instances in the execution since the data are permuted for each execution.

Table 5: Performance metrics for SVM & KNN (noisy data)

Classifier	Sensitivity	Specificity	Accuracy	K-Fold Loss
SVM-L	30%	60%	75%	0.0222
SVM-R	90%	92%	91.5%	0.0286
KNN	86%	87.2%	89.7%	0.0476

V. CONCLUSION

In the proposed work, microscopic blood smear images have been analyzed and detection of ALL is made using morphological features of blast cells. These extracted features were used to discriminate malignant cells from benign cells and the classification is conducted by SVM and KNN classifier. The results obtained indicates KNN yields better results when compared to other classifiers only when noise free data is used. But when noisy data is given, SVM-R yields better results when compared to KNN [18].

The future work is to develop a classification system that classifies the malignant (ALL) cells to its subtypes-L1, L2 and L3 [19]. These subtypes can also be classified by using SVM or KNN. Since SVM is a binary classifier, it is important to convert binary SVM into a multi class SVM. This can be done by using Error Correction Output Codes (ECOC) [20] [21]. KNN can also be used for this subtype classification by adjusting the value of K [22]. This can also be implemented by using neural networks for better accuracy [23]. The complexity lies in differentiating these subtypes. Hence more features are needed to be extracted from both nucleus and cytoplasm [24]. Since death rate due to this disease [25] is increasing rapidly, the early detection of this disease is necessary.

Acknowledgment

We like to express our gratefulness to **N. Shaikh Mohammad Bilal, M.E., Assistant Professor,** Department of Computer Science, KJ Somaiya College, Vidyavihar and Mumbai for providing the data sets and guidance throughout the project work.

Bibliography

[1]. https://www.cancercenter.com

[2]. http://www.hematology.org/Patients/Cancers/Leukemia.aspx.

[3]. Lorentzo Putzu, Giovanni Caocci, Cecilia Di Ruberto "Leucocyte classification for Leukemia detection using image processing techniques". Artificial Intelligence in Medicine, pages 179-191 (2014).

[4].Endah Purwanti and Evelyn Calista , "Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features", IOP Conf. Series: Journal of Physics: Conf. Series 853 (2017).

[5]. M. Bennet Rajesh and S. Sathiamoorthy, "Classification of Leukemia Image Using Genetic Based K-Nearest Neighbor (G-KNN)", Asian Journal of Computer Science and Technology, Volume 7 No.2, pages.113-117 (2018).

[6]. Jyoti Rawat, Annapurna Singh, H.S. Bhadauria, Jitendra Virmani, Jagtar Singh Devgun, "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia", Bio cybernetics and Biomedical Engineering Journal, Volume 37, Issue 4, July 2017, Pages 637-654.

[7]. Ruggero Donida Labiti, Vincenzo Piuri, Fabio Scotti "ALL-IDB Acute Lymphoblastic Leukemia Image Database for image processing". IEEE Transactions, Pg. No. 2045-2048, (2011).

[8].Himali Vaghela, Hardik Modi, Manoj Pandiya, Potdar M.B, "A Novel Approach to detect chronic leukemia using shape based feature extraction and identification with digital image processing. International Journal Applied Information Systems (IJAIS). Volume 11, No.5 (2016).

[9]. Himali Vaghela, Hardik Modi, Manoj Pandiya, Potdar M.B "Leukemia Detection using Digital Image Processing Techniques". International Journal of Applied Information Systems (IJAIS), Volume 10-No.1 (2015).

[10]. Rajesh Kumar, Rajeev Srivastava and Subodh Srivastava, "Detection and Classification of Cancer from Microscopic Biopsy images using clinically significant and biologically interpretable features". Hindawi publishing corporation, Journal of Medical Engineering, Volume 2015, Article ID 457906, 14 pages (2015).

[11]. Lim Huey Nee, Mohd Yusuff Mashor, Rosline Hassen,"White Blood Cell Segmentation for Acute Leukemia Bone Marrow images". Journal of Medical Imaging and Health Infomatics. Volume 2, Pg. No: 278-28 (2012).

[12]. Pratik M. Gumble and Dr.S.V.Rode, "Analysis & Classification of Acute Lymphoblastic Leukemia using KNN Algorithm", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 5 Issue: 2 pages 94 – 98 (2017).

[13].https://in.mathworks.com/help/matlab/learn_matlab/product-description.html

[14]. Vasuki Shankar et al, "Automatic detection of Acute Lymphoblasitc Leukemia using Image Processing", IEEE International Conference on Advances in Computer Applications (ICACA) (2016).

[15]. Rama Gaur, V.S. Chouhan,"Classifiers in Image Processing". International Journal on Future Revolution in Computer Science & Communication Engineering. Volume: 3 Issue: 6 .page n0.22 – 24, (2017).

[16].https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

[17]. Ahmed S. Negm, Osama A. Hassan, Ahmed H. Kandil, "A decision support system for Acute Leukaemia classification based on digital microscopic images", Alexandria Engineering Journal, volume 57, Issue 4, Dec 2018, Pages 2319-2332.

[18]. https://www.cancer.gov/types/leukemia

[19].https://www.mayoclinic.org/diseases-conditions/acute-lymphocytic-leukemia/symptoms-causes/syc-20369077

[20]. Kumar PS and Vasuki S, "Automated diagnosis of acute lymphocytic leukemia and acute myeloid leukemia using multi-SV", J Biomedical Image Bioengineering, (2017). Volume 1 Issue 1.

[21].https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

[22].https://www.verywellhealth.com/understanding-white-blood-cells-and-counts-2249217

[23]. V.Venmathi. K.N.Shobana, Akshaya Kumar, D.Gajesh Waran (2017). "Leukemia detection using image processing". International Journal for Scientific Research & Development (IJSRD) Vol. 5, Issue 01, 2017

[24]. Sarmad Shafique and Samabia Tehsin, "Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks", Technology in Cancer Research & Treatment (2018), Volume 17, pages 1-7.

[25]. https://www.medicalnewstoday.com/articles/142595.php