Grenze International Journal of Engineering and Technology, Special issue

Detection of Acute Lymphocytic Leukemia using Blast Cell Morphological Features

S. Hariprasath[1], T.Dharani[2] and Dr.M.Santhi[3]
[1]Saranathan College of Engineering /Department of ECE, Trichy, India
Email: hariprasath-ece@saranathan.ac.in
[2-3] Saranathan College of Engineering /Department of ECE, Trichy, India
Email: dharanithirugnanam@gmail.com, santhim-ece@saranathan.ac.in

***Abstract*—Precise detection of leukemia/ blood cancer is one of the major issues in the medical field so that the victims can be treated effectively. 'Leukemia' affects the lymphatic system of our body. Potent infection fighter cells that are white blood cells are involved in causing leukemia. If leukemia causes abnormalities in lymphocytes, it is known as lymphoblastic/ lymphocytic leukemia. If leukemia causes abnormalities in myelocytes, it is known as myeloblastic/ myelocytic leukemia. Leukemia can also be further classified as 'Acute' and 'Chronic'. Acute leukemia is more decisive than chronic leukemia since it causes immediate death. In this paper, the detection and classification of Acute Lymphocytic Leukemia (ALL) and their subtypes are focused. Supervised classifiers such as SVM and KNN classifiers are evaluated for the detection and classification of cells. From the results obtained, it is determined that the KNN classifier is more suitable for classification when noise-free data is used.**

***Index Terms*— Leukemia, ALL, Pattern Classification system, SVM, KNN.**

I. INTRODUCTION

The involvement of Image processing and machine learning techniques in the medical field is through the digitalized medical images that can be analyzed with the help of a computer. The growth of research and development in image analysis has contributed to the medical field. Medical images are considered as an important tool that is used for the diagnosis and analysis of many diseases. Such images can also be used for research and teaching purposes. The digital medical images that are used in this work are microscopic Peripheral Blood (PB) smear images.

This work is an attempt to apply digital image processing and ML techniques in the area of medical image analysis and recognition, in particular, Leukemia. The focus of this work is on developing a methodology to diagnose and classify acute lymphocytic leukemia, based on cell morphology, into its subtypes.

Leucocytes that are White Blood Cells (WBC) has an important part in providing immune to our body. Blood Cancer/ Leukemia is one of the lethal diseases in India as the death rate due to leukemia is growing every year. When a person is affected by leukemia, the morphology of the cells gets modified. Hence these cells can be used to discriminate healthy and leukemic cells [1].

The FAB (French-American-British) classification system is the generally used cancer classification system. This system classifies 'Acute' leukemia into Acute Lymphocytic Leukemia (ALL) and Acute Myelocytic Leukemia (AML). It also classifies ALL into its subtypes such as L1, L2 and L3. And classifies AML into its

Grenze ID: 01.GIJET.5.1.22

subtypes such as M1 to M7. These subtypes can be recognized by their nuclear structure, cytoplasmic structure and other morphological features. [2].
When compared to the other types of leukemia, children are more affected by ALL type of leukemia. Thus the precise diagnosis of this disease is important for treating the affected person effectively.
In this paper, ALL type of leukemia is diagnosed using image processing techniques. The cell structure of malignant lymphocyte is called 'lymphoblast' where the shape of the nucleus may be irregular, the overlap may occur with the adjacent cells and there is an abnormality in the nucleus and cytoplasmic ratio. Lymphoblasts can also refer to immature cells that typically differentiate form mature lymphocytes. The blood of the healthy and leukemic person is given in figure 1.

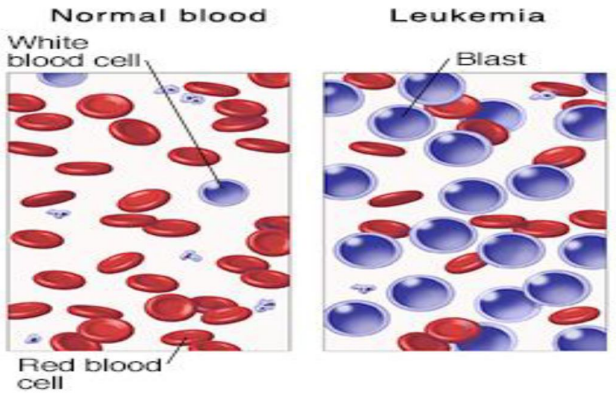

Figure 1. Difference between healthy and cancerous blood

II. EXISTING SYSTEMS

Several image processing methods have been incorporated for obtaining the precise detection of diseases. The accuracy of these methods depends on many factors such as features used, classifiers used and quality of the dataset used.
Lorentzo Putzu et al have designed a detection system that detects and classifies the white blood cells. They analyzed 108 images from ALL-IDB dataset which is a commonly used standard cancer dataset. Leucocyte identification is made possible by using 131 features (shape, color and texture descriptors). Various SVM classifiers such as SVM-L, SVM-Q, SVM-P, and SVM-R are analyzed and the results are compared with KNN classifiers and Naive Bayes (NB) classifier and decision tree classifiers. From the results, it is concluded that SVM-R achieved better accuracy at different lighting conditions [3].
Endah Purwanti and Evelyn Calista have proposed a model that uses the combination of shape and histogram features for ALL detection [4]. KNN classifier is used for classifying the images from ALL-IDB into healthy and malignant types. Choosing the K value is a vital issue in the KNN classifier. Various K values are analyzed for classification and better results are obtained when the K value is taken as 7. [4]
Bennet Rajesh and S. Sathiamoorthy have combined genetic algorithm and KNN for removing noise and preprocessing the images. A Median filter is used to reduce the salt and pepper noise from the image. To find the nearest neighbor, 'Euclidean Metric' is used and the results produced better accuracy when compared with SVM and Multilayer Perceptron Neural Network (MLPNN) [5].
Generally used cancer database for detection and classification is ALL-IDB. In 'ALL-IDB: the acute lymphoblastic leukemia image database for image processing' by Ruggero donida labati, vincenzo piuri, fabio scotti, an automated method for detection of lymphoblast is done in the sequence of steps such as segmentation, identification of white cells, identification of lymphocytes. The database used is ALL-IDB that has two distinct versions such as (ALL-IDB1 & ALL-IDB2) [6].

III. METHODOLOGY

The major steps involved in the automated detection of any disease are:

- Image Acquisition
- Image pre-processing

- Image Segmentation
- Feature extraction
- Classification

A. Image Acquisition

The first step in the process of detection of a disease is to gather standard images. ALL-IDB is the most commonly used cancer dataset. It consists of 100 images that are well captured by a standard camera like Canon Power Shot G5 camera [6]. The sample images of healthy and leukemia affected leucocytes are given in figure 2.

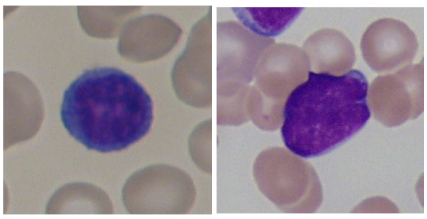

Figure 2. Samples from ALL-IDB for healthy and malignant cells

B. Image Pre-processing

The given sample images are in the RGB format which cannot be used to identify WBCs. Hence it is converted into CMYK format since all other components have a yellow component except the leucocytes.

C. Image Segmentation

A Pre-processed image is needed to be enhanced and segmented in order to obtain the region of interest. In this work, enhancement is done by equalizing the image intensities by using histogram equalization. Zack thresholding or triangle method is used to fix the threshold. It is applied to the image histogram which results in a straight line that connects the maximum and minimum histogram value. The resultant fixed threshold value is 0.5664.

D. Feature Extraction

Since discrimination between healthy and malignant cells is based on the nuclear and cytoplasmic structure, features must be extracted from the cell's nucleus and cytoplasm. Initially, binary image is obtained from the green component of RGB format and it is combined with a binary image obtained from the a* component (CIE Lab Colour Space) by a thresholding operation. To obtain the cytoplasm of the cell, a binary image holding the whole leucocyte is subtracted from the image holding only the nucleus. Then, morphological features like statistical and textural descriptors are extracted from the nucleus and cytoplasm. Shape descriptors [7] and colour descriptors are collectively known as Statistical descriptors. Totally 35 features are extracted from the image. The extracted features comprise of 9 shape descriptors, 6 colour descriptors and 20 textural descriptors [8] [9]. Table 3 shows the texture features of healthy and cancerous cells.

E. Classifier System

The given testing samples can be classified into healthy and cancerous classes with the help of a supervised classifier. In this paper, KNN classifiers are used and compared with the SVM classifier [10].

F. ALL Subtypes

The major types of ALL are L1, L2 and L3. They are explained below and the sample of all subtypes are shown in figure 4, 5 and 6.

ALL-L1 type: The lymphoblast cells will be small and the nucleus will have a regular shape, little cytoplasm, and high nucleic cytoplasmic ratio.

ALL-L2 type: L2 type cells will have variable irregular nuclear structure, variable nucleic cytoplasmic ratio. The size of the cell is always larger than L1 type cells. Hence L2 cells can be easily identified by heterogeneous nuclear structure.

ALL-L3 type: L3 type cells will have vacuoles in the cell structure. The nucleus of L3 is in round/ oval structure. The nuclear structure of L3 is always larger than the L1 type nucleus.

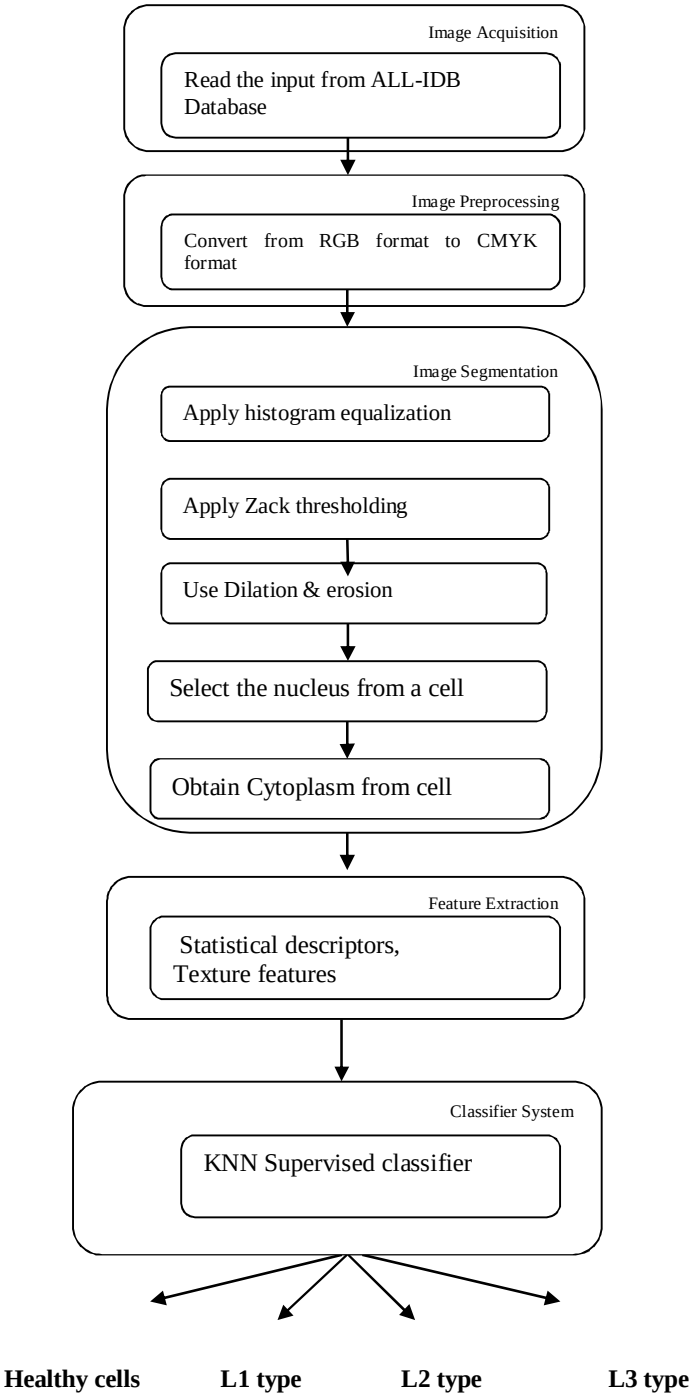

Figure 3. Proposed Methodology

K-NN Classifier

The K-NN is a supervised machine learning classifier. It produces results based on the majority voting technique. Choosing the value of K in KNN is always a major issue. The value of K depends on the no. of training samples used. K is always chosen as an odd number. The nearest neighbor can be detected by using any distance metric such as Euclidean distance, Mahalanobis distance, Minkowski distance, etc. In this work,

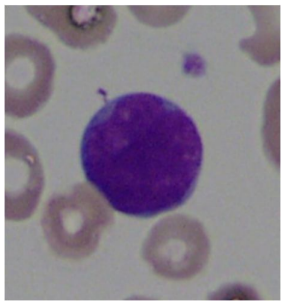

Figure 4. ALL-L1 sample cell

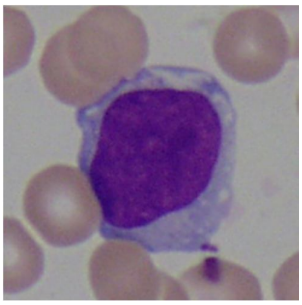

Figure 5. ALL-L2 sample cell

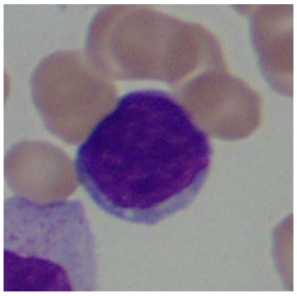

Figure 6. ALL-L3 sample cell

'Euclidean distance metric' is used for determining the nearest neighbor [11]. Figure 7 shows the KNN model for 2 class problem.

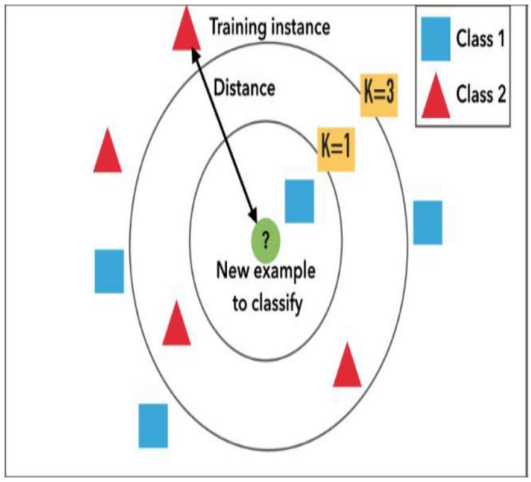

Figure 7. KNN model for two class problem

IV. RESULTS & ANALYSIS

SVM classifier classifies the test samples into two classes: healthy and malignant (cancerous) cells. Initially, SVM selects the best parameters and the model is tested and visualized [12]. Test points that are very close to the hyperplane are chosen as 'support vectors'. In this work, SVM selects solidity and a standard deviation is taken as the best parameters for locating an optimal hyperplane [13].

In KNN, 5 nearest neighbors & K value is fixed at 5 in our execution. After dividing the test samples into two classes, healthy (benign) and cancerous cells, the ALL cells are needed to be classified into three classes that is L1, L2 and L3 [14]. The labeling of these cells can be done by using a linear classifier. The labels of the subtype

cells are given as Healthy-1, L1 type-2, L2 type-3, L3 type-4 [15] [16]. The Cancer dataset ALL-IDB (100 samples) contains 35 healthy cells and 65 malignant cells (L1-37 samples, L2-9 samples, L3-19 samples).
To calculate the performance metrics listed in table 2, the confusion/ error matrix is determined. Confusion matrix / Error matrix is a table that is used to evaluate the performance of a classification model. It is generally used for a supervised classifier.
To compute the performance metrics like accuracy, sensitivity and specificity, values of TP, FP, TN, and FN are needed.
True positive (TP) = the number of cases properly recognized as malignant.
False positive (FP) = the number of cases improperly recognized as malignant.
True negative (TN) = the number of cases properly recognized as benign.
False negative (FN) = the number of cases improperly recognized as benign.

TABLE I: CONFUSION MATRIX FOR 4 CLASSES

		PREDICTED			
		Healthy (A)	L1 type (B)	L2 type (C)	L3 type (D)
ACTUAL	Healthy (A)	TPA	WAB	WAC	WAD
	L1 type (B)	WBA	TPB	WBC	WBD
	L2 type (C)	WCA	WCB	TPC	WCD
	L3 type (D)	WDA	WDB	WDC	TPD

Where TP X= True Positive for X, Wxy = Classification Error (True Negative, False Positive, False Negative).

TABLE II. PERFORMANCE METRICS AND MATHEMATICAL FORMULA

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

TABLE III. TEXTURE FEATURES FOR HEALTHY AND CANCEROUS SAMPLES

Type of cell	**Auto-Correlation**	**Contrast**	**Energy**	**Entropy**
Benign	19.14	0.0301	0.4217	1.180
Malignant	25.62	0.0583	0.2555	1.605

Figure 8 & 9 shows the implementation results of SVM-L & SVM-R respectively. The results displays the actual and predicted labels and the performance metrics obtained.

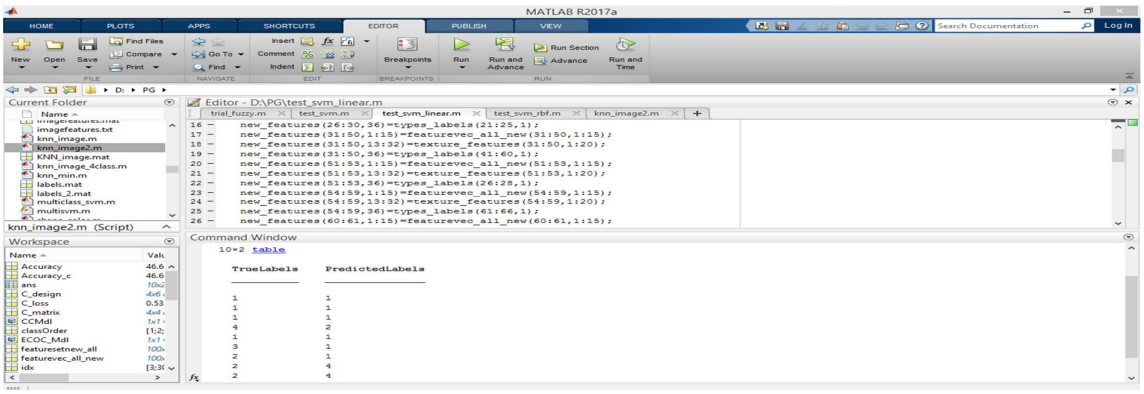

Figure 8. The output of SVM-L showing the actual and predicted labels

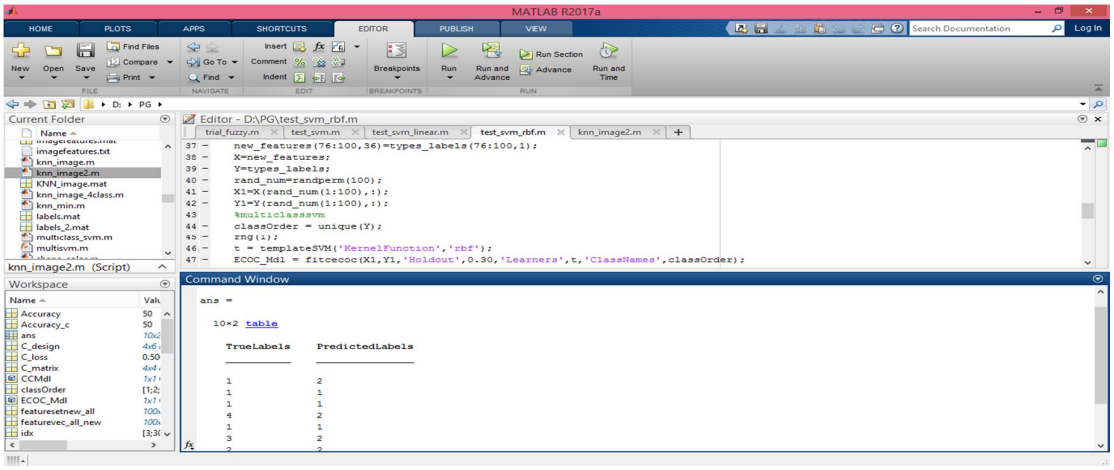

Figure 9. The output of SVM-R showing the actual and predicted labels

From the SVM results obtained, it can be seen that SVM-R produces better results when compared to SVM-L classifier.

Figure 10 shows the results obtained for KNN classifier (4 class). Since the dataset is permuted for each execution, the results given in the table are the average results for three instances.

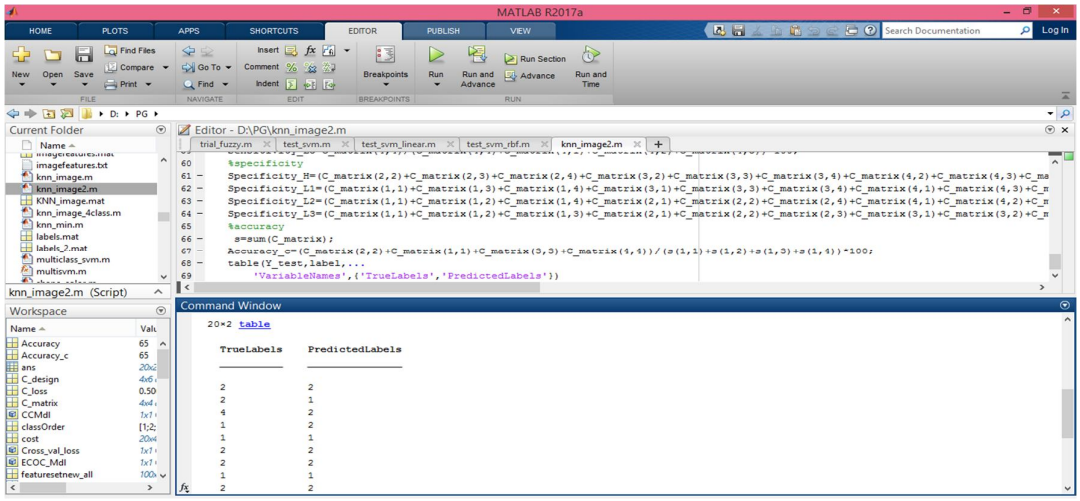

Figure 10. The output of KNN showing the actual and predicted labels

The results obtained for all the classifiers are given table 4 & 5. To analyse the classifier efficiently, the sample points are permuted for each execution.

TABLE IV: SENSITIVITY OBTAINED FOR SVM AND KNN CLASSIFIERS

Method	Sensitivity of H %	Sensitivity of L1 %	Sensitivity of L2 %	Sensitivity of L3 %
Dataset (not permuted)	100	90.9	0	16.66
Dataset permuted	100	72.72	25	16.66
SVM-L	90	27.27	0	33.33
SVM-R	50	90.9	0	0
KNN	80.5	83.76	0	25

TABLE V: SPECIFICITY OBTAINED FOR SVM AND KNN CLASSIFIER

Method	Specificity of H %	Specificity of L1 %	Specificity of L2 %	Specificity of L3 %
Dataset (not permuted)	100	57.8	100	95.8
Dataset permuted	95	78.94	81.48	100
SVM-L	60	73.68	100	87.5
SVM-R	95	26.3	100	100
KNN	94.82	48.73	100	95.83

From the results (shown in figure 11) it is concluded that KNN is well suitable for the classification of healthy and cancerous cells and its subtypes, only when small and noise-free data set is used.

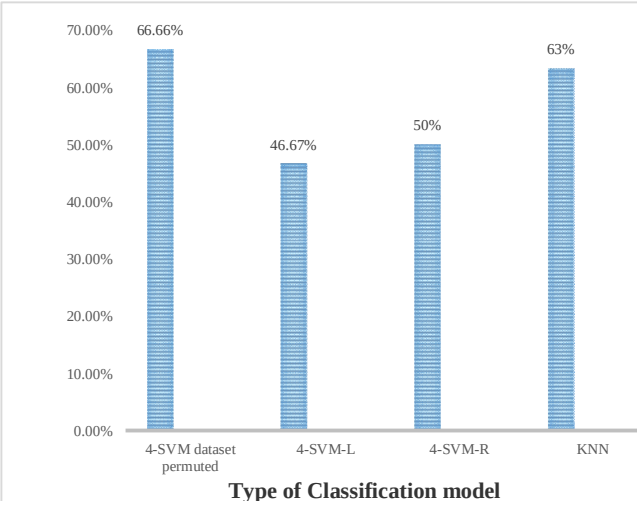

Figure 11. Classification Accuracy chart

V. Conclusions

In the proposed method, peripheral blood smear samples have been analyzed and the detection of Acute Lymphocytic Leukemia (ALL) is done [15]. The detection of this disease is made possible with the help of blast cell morphological features. These features are used to discriminate cancerous cells from healthy cells and the classification is performed by SVM and KNN classifier. The results obtained show that the KNN classifier produces better results when noise-free data is used. The future work of this method is to design and develop a fuzzy based classification system that classifies ALL cells into its subtypes- L1, L2 and L3 cells [16] [17] with better accuracy. This can be made conceivable by utilizing more features from the cell's nucleus and cytoplasm.

Acknowledgment

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

References

[1] https://www.mayoclinic.org/diseases-conditions/acute-lymphocytic-leukemia/symptoms-causes/syc-20369077
[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] Ruggero Donida Labiti, Vincenzo Piuri, Fabio Scotti "ALL-IDB Acute Lymphoblastic Leukemia Image Database for image processing". IEEE Transactions, Pg. No. 2045-2048, (2011).
[7] 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).
[8] 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).
[9] 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).
[10] 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).
[11] 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).
[12] 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.
[13] 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.
[14] https://www.verywellhealth.com/understanding-white-blood-cells-and-counts-2249217
[15] https://www.cancercenter.com
[16] https://www.cancer.gov/types/leukemia
[17] https://www.merriamwebster.com/dictionary/lymphocytic%20leukemia