INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

**ISSN: 2454-132X**
**Impact factor: 4.295**
***(Volume 3, Issue 6)***
Available online at www.ijariit.com

# Detection and Classification of Blood Cancer from Microscopic Cell Images Using SVM KNN and NN Classifier

**Sachin Paswan**
*Student*
*Raipur Institute of Technology, Raipur, Chhattisgarh*
*Sksachin0@gmail.com*

**Yogesh Kumar Rathore**
*Assistant Professor*
*Raipur Institute of Technology, Raipur, Chhattisgarh*
*yogeshrathore23@gmail.com*

***Abstract: Leukemia is a cancer of the blood and bone marrow, the spongy tissue confidential the bones where blood cells are made. Acute myeloid leukemia (AML) is one of the most common types of leukemia among adults. The signs and symptoms of leukemia are non-specific in nature and also they are comparable to the symptoms of other mutual disorders. Manual microscopic inspection of stained blood smear or bone marrow aspirate is the only way to an effective diagnosis of leukemia. But this method is time consuming and less accurate. In this paper, a technique for automatic detection and classification of AML in blood smear is presented. K-means algorithm is used for segmentation. KNN, NN, and SVM are used for classification. GLCM is used for optimizing the spectral features. The local binary pattern is used for texture description. Blood microscope images were tested and the performance of the classifier was analyzed.***

***Keywords: Automated Leukemia detection, Acute Lymphoblastic Leukemia, Lymphocyte Image Segmentation, Machine Learning.***

## I. INTRODUCTION

Microscopic analysis of peripheral blood smear is a critical step in the detection of leukemia. However, this type of light microscopic assessment is time consuming, inherently subjective, and is governed by hematopathologists clinical acumen and experience. To circumvent such problems, an efficient computer aided methodology for quantitative analysis of peripheral blood samples is required to be developed. In this paper, efforts are therefore made to devise methodologies for automated detection and sub-classification of Acute Lymphoblastic Leukemia (ALL) using image processing and machine learning methods. Choice of appropriate segmentation scheme plays a vital role in the automated disease recognition process. Accordingly to segment the normal mature lymphocyte and malignant lymphoblast images into constituent morphological regions novel schemes have been proposed. In order to make the proposed schemes viable from a practical and real–time stand point, the segmentation problem is addressed in the supervised framework. These proposed methods are based on neural network, feature space clustering, and SVM field modeling, where the segmentation problem is formulated as pixel classification, pixel clustering, and pixel labeling problem respectively. A comprehensive validation analysis is presented to evaluate the performance of three proposed lymphocyte image segmentation schemes against manual segmentation results provided by a panel of hematopathologists. It is observed that morphological components of normal and malignant lymphocytes differ significantly. To automatically recognize lymphoblasts and detect ALL in peripheral blood samples, an efficient methodology is proposed. Morphological, textural and color features are extracted from the segmented nucleus and cytoplasm regions of the lymphocyte images. An ensemble of classifiers represented as SVM comprising of three classifiers shows highest classification accuracy of 83.33% in comparison to individual members.

These methods include lymphoblast image segmentation, nucleus and cytoplasm feature extraction, and efficient classification. To subtype leukemia blast images based on cell lineages, an improved scheme is also proposed and the results are correlated with that of the flow cytometer. Using this scheme the origin of blast cells i.e. lymphoid or myeloid can be determined. An ensemble of decision trees is used to map the extracted features of the leukemic blast images into one of the two groups. Each model is studied

 

separately and experiments are conducted to evaluate their performances. Performance measures i.e. accuracy efficacy of the proposed automated systems with that of standard diagnostic procedures.

### A. LEUKEMIA

Leukemia is a group of heterogeneous blood-related cancers, differing in its aetiology, pathogenesis, prognosis, and response to treatment (Bain, 2010). Leukemia is considered as a serious issue in modern society, as it affects both children and adults and even sometimes infants under the age of 12 months. In children, leukemia is considered as the most common type of cancer, while, in adults, the World Health Organization report shows that leukemia is one of the top 15 most common types of cancer (Kampen, 2012). To better understand leukemia, the next sections are dedicated to the discussion of the blood cells lineage, types of leukemia, diagnostic methods currently in use, treatments options as well as prognostic factors.

### B. TYPES OF LEUKEMIA

Lab tests help the doctor find out the type of leukemia that you have. For each type of leukemia, the treatment plan is different.
Acute and Chronic Leukemias
Leukemias are named for how quickly the disease develops and gets worse:

 **Acute:** Acute leukemia usually develops quickly. The number of leukemia cells increases rapidly, and these abnormal cells don't do the work of normal white blood cells. A bone marrow test may show a high level of leukemia cells and low levels of normal blood cells. People with acute leukemia may feel very tired, bruise easily, and get infections often.

 **Chronic:** Chronic leukemia usually develops slowly. The leukemia cells work almost as well as normal white blood cells. People may not feel sick at first, and the first sign of illness may be abnormal results on a routine blood test. For example, the blood test may show a high level of leukemia cells. If not treated, the leukemia cells may later crowd out normal blood cells.

### C. Myeloid and Lymphoid Leukemia's

Leukemia's are also named for the type of white blood cell that is affected:

 **Myeloid:** Leukemia that starts in myeloid cells is called myeloid, myelogenous, or myeloblastic leukemia.

 **Lymphoid:** Leukemia that starts in lymphoid cells is called lymphoid, lymphoblastic, or lymphocytic leukemia. Lymphoid leukemia cells may collect in the lymph nodes, which become swollen.

### D. Four Most Common Types of Leukemia

 **Acute myeloid leukemia (AML)** affects myeloid cells and grows quickly. Leukemic blast cells collect in the bone marrow and blood.

About 15,000 Americans will be diagnosed with AML in 2013. Most (about 8,000) will be 65 or older, and about 870 children and teens will get this disease.

 **Acute lymphoblastic leukemia (ALL)** affects lymphoid cells and grows quickly. Leukemic blast cells usually collect in the bone marrow and blood [7].

More than 6,000 Americans will be diagnosed with ALL in 2013. Most (more than 3,600) will be children and teens.

 **Chronic myeloid leukemia (CML)** affects myeloid cells and usually grows slowly at first. Blood tests show an increase in the number of white blood cells. There may be a small number of leukemic blast cells in the bone marrow. About 6,000 Americans will be diagnosed with CML in 2013. Almost half (about 2,900) will be 65 or older, and only about 170 children and teens will get this disease.

 **Chronic lymphocytic leukemia (CLL)** affects lymphoid cells and usually grows slowly. Blood tests show an increase in the number of white blood cells. The abnormal cells work almost as well as the normal WBC. About 16,000 Americans will be diagnosed with CLL in 2013. Most (about 10,700) will be 65 or older. This disease almost never affects children or teens. Other, less common types of leukemia will account for more than 6,000 new cases in 2013.

 

## II. LITERATURE REVIEW

In [9] an iterative thresholding algorithm is used for segmentation purpose especially from noisy images. This algorithm overcomes the problem of cell extraction and segmentation from heavy noisy images. This algorithm works over the adjusted threshold of images iteratively providing robustness to the image.

In [10] discusses the malarial image processing system. This system detects and classifies malaria parasites in Giemsa stained blood slides images. Then after parasitaemia evaluation is done. A morphological approach to cell image segmentation is more precise than the classical watershed-based algorithm is shown in this paper. Grey scale granulometries are applied based on opening with disk-shaped elements, flat and non-flat. Non flat disk shaped structuring element enhances the roundness and the red cells compactness.

In [11] a system classifies and identify malaria parasite by using microscopic images of blood cells. Morphological approach and the major necessities in developing this system are the best techniques for blood cell images segmentation.

In [12] research work on an Automated Cell Count method is described. A precise method of segmentation for counting white blood cells automatically is presented here. First, a simple thresholding approach is applied and the algorithm is derived from blood smear images from a priori information. The labels are adjusted then in order to produce meaningful results. This approach uses knowledge of the blood cell structure. This method is more influential as compared to traditional methods which use information of local context. It can perform accurate segmentation of white blood cells though they have un-sharp boundaries.

In [13] using a filter bank of a' trous wavelet filters, curvelet transform implements curvelet sub-bands and uses a ridgelet transform as a component step, and idea throughout is that transforms should be over complete, more willingly than critically sampled. In this digital transforms are applied for de-noising of some standard images rooted in white noise. A combination of geometric distance and an enhanced distance transform combining intensity gradients is used for the watershed step in [14].An explicit mathematical model for characteristics of cell nuclei like size and shape measures is included. For each detected nucleus, a confidence score is computed by measuring the suitability of nucleus in the model.

Paper [15] shows the usefulness of an automatic morphological method to recognize the Acute Lymphocytic Leukemia (ALL) with the help of images of peripheral blood microscope. The presented methodology individuates the leucocytes from the others blood cells, after that, it selects the lymphocyte cells (the cells causes acute leukemia), morphological indexes from those cells are evaluated then after and at last classification is performed whether the presence of the leukemia is there or not.

## III. SYSTEM MODEL

A microscope has been implemented by us for acquiring the blood cell images and has implemented MATLAB software R2015b for image processing. 100 microscopic blood cell images of various sizes have been acquired for testing. Our propounded approach ensures efficacious classification of blood microscopic image with 83.33 % accuracy. The segmentation is carried out in the following steps which are mentioned below:

### *A.* Acquisition of Images

By the aid of a microscope the blood images are acquired in this image are displayed in a 2D matrix, where the pixels of the image are imagined as the element such matrices are entirely dependent on the field of view and matrix size. Our propounded approach implements MATLAB for storing the images in a database and is displayed in Lab scale dimension of 512 *512.

### B. Pre-processing

In this specific phase, the input image is compounded in a way such that even infinitesimal detail are ameliorated and unnoted noises are filtered commonly used noise filter methodologies are implemented which aids in the procurement of the feasible results. Conspicuous edges, sharpened image, and noise reduction are the outcomes of image enhancement. Enhancement reduces the blotting out of the image and thus reduces the chances of getting twisted results from the intervening system. Eventually, segmentation is also applied. This improved and fine-tunes image helps in edge determination and ameliorates the overall quality of the image.

The blood microscopic images which are procured are stored in a MATLAB database and eventually transposed to LAB scale image having a dimension of 512*512; 2) further the image is processed to remove any unwanted presence of noises.) Eventually, the high pass filter (i.e., Gaussian Filter) works upon the refined image, which is of a higher resolution, aids in the procurement of sharpened image and also aids in edge detection.

 

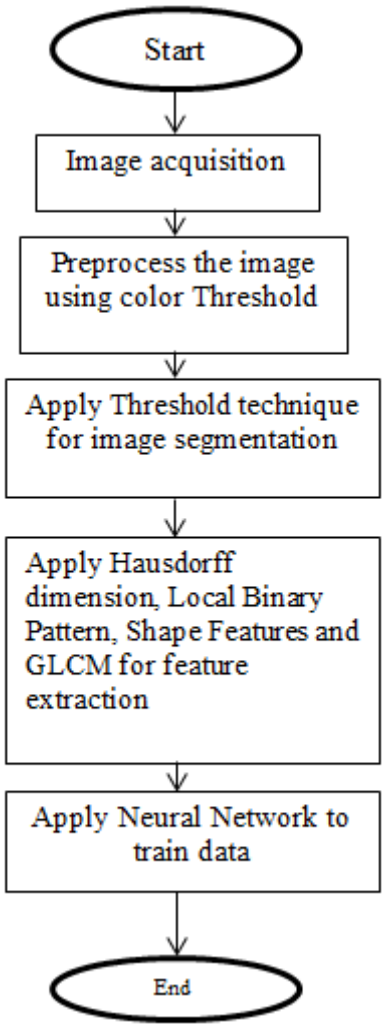


**Figure 1: Flow diagram of work**

### *C.* Segmentation

The technique of partitioning the image into segment can be defined as image segmentation. Considering the similar property, segmentation is implemented. This similar property is cluster together our propounded approach implements Lloyd's clustering technique which aids in the segmentation of blood microscopic images on the basis of alike properties. This technique broadens the k-mean clustering algorithm by introducing repeated segmentation scheme which explores the centroid of each set in the segment and eventually re-segment the input based on the closest centroid. This technique aids in the extraction of important image characteristics, based on which information can be easily perceived.

A simple thresholding approach is applied to give initial labels to pixels in the blood cell images. The algorithm is based on a priori information about blood smear images. Then the labels are adjusted with a shape detection method based on large regional context information to produce meaningful results.

### *D.* Feature Extraction

The process of defining a large set of surplus data into a feature set with a reduced dimension is known as feature extraction. The performance of the classifier is greatly ameliorated by the feature selection; hence the correct feature selection becomes an important step [13]. The principle concern in generating the blood cell features is that the recognition of the various blast cells with the highest accuracy is a great deal. In feature extraction, the input data are transformed into a set of features. The classifier performance is greatly influenced by the feature selection and therefore, the correct selection of features is a very crucial step. The features considered for calculation are Hausdorff dimension, Shape features, Texture features [16] and GLCM (Grey Level Co-occurrence Matrix) features. Hausdorff dimension is the ratio of the log of a number of squares in the superimposed grid to the number of occupied squares.

The following features were implemented by us, which corresponds to the lymphocyte's nucleus and myelocyte's nucleus:

#### a. Color Features

In this scheme, the color spaces corresponding to the RGB scale will be converted into HSV scale. Mean color values will be procured.

 

**b. Texture Features**

It comprises of the following: i) energy, ii) homogeneity, iii) entropy, iv) angular second momentum, v) correlation, vi) contrast, are procured.

**c. GLCM feature**

The GLCM feature calculation is an image analysis technique [12]. Different texture features like energy, contrast, entropy, correlation are extracted using this method.

**d. Hausdorff dimension**

For various quantitative measurements, fractals have been used for a long time. Fractal dimension is a statistical quantity that indicates how completely a fractal appears to fill space. Haushorff dimension and the packing dimension are the most important theoretical fractal dimensions. In practical applications, the box-counting dimension is widely used, due to the fact that it is easy to implement. In box counting algorithms, the number of boxes covering a point set is a power law function of the box size. All fractal dimensions are estimated as the exponent of such power laws and are real numbers that characterize the fractalness (texture or roughness) of the objects. The perimeter roughness of the nucleuscan is used to differentiate myeloblasts.

**Steps to find Hausdorff dimension:**

1) Obtain the binary image from the colour image

2) The nucleus boundaries are traced out using edge detection technique

3) The edges are superimposed by a grid of squares.

4) Then, the hausdorff dimension is defined as

$$HD = \frac{\log(R)}{\log(\mathrm{R}(s))}$$

Where R is the number of squares in the superimposed grid and R(s) is the number of occupied squares.

**E. Classification Stage**

In this particular stage, the principle concern is on the selection strategies of relevant classification technique, as it is the intriguing problem because if an appropriate choice is given, the data available can ameliorate the accuracy in the scoring of credit [14], [15]. Numerous statistical methods, which mainly focus on solving binary classification problems, are widely available.
In this specific paper, we have implemented a Neural Network (NN), because the data is trained hastily and aids in data classification.

## IV. EXPERIMENTAL PROPOSED METHOD

**A. KNN**

The KNN binary (as two class) is given more accurate data classification which beneficial to select k as an odd number which avoids the irregular data. The KNN procedure is the technique in ML procedures: It is an object which classified through a mainstream selection of its neighbors, with the determination assigned occurrence for most mutual class amongst its *k* nearest neighbors (*k* is a positive integer, classically small). Classically Euclidean distance is used as the distance metric; however, this is only suitable for endless variables. In such situation as the classification of text, alternative metric, intersection metric or Hamming distance can be used.
KNN is a new process that deliveries all available cases and categorizes novel cases built on an evaluation quantity (e.g., distance functions).

KNN procedure is identical simple. It works built on a minimum distance from the interrogation instance to the training samples to regulate the K-nearest neighbors. The information for KNN procedure contains numerous attribute which will be used to categorize. The information of KNN can be any dimension scale from insignificant, to measurable scale.

**B. NN (**Training and Testing Procedure**)**

By using feed forward neural network and back propagation algorithm, width, position and the average intensity of chromosome were determined back propagation algorithm achieves high accuracy with minimum training time, which makes it suitable for real-time chromosome classification in the laboratory.in our paper, segmentation is done by using image processing and classification is done by using feed forward neural network and back propagation algorithm. In this paper, feed forward neural network & back propagation algorithm was applied to the classification of images by considering intensity inhomogeneity, which often exists in the images. The basic architecture consists of three types of neuron layers: input, hidden, and output. In feed-forward networks, the signal flow is from input to output units, strictly in a feedforward direction. The data processing can extend over multiple layers of units, but no feedback connections are present

 

### C. SVM

Since SVM is inherently a binary classification system, there are numerous methods available to extend it to a multiclass classifier. SVM is inherently a binary classifier, thus it can classify input into either of two classes for which it has been trained. To use SVM for the multiclass problem, a number of approaches have been employed over the years [2, 7] that use a combination of several binary SVM classifiers. Some of the popular methods are: a one-versus-all method using winner takes all strategy (WTA_SVM), one-versus-one using maxwins voting method (MWV_SVM). A good general strategy called pairwise coupling using posterior probabilities (PWC_PSVM) was also proposed by Hastie and Tibshirani [8]. In the first approach "one against all", the test data object is classified based on the greatest value that is determined. For N number of classes N numbers of SVMs are used which generates N decision functions. Even though this method is faster than other methods, it suffers from inconsistent training sets. By the second approach "one against one", between each pair of classes an SVM is generated. Then, the max-win voting will decide to which class the object belongs. The third method "pairwise coupling" strategy uses the combined probability output of all the one-versus-one methods and generates the posterior probabilities pi. After all the estimates calculated, this method will assign the test vector to that class which has largest posterior probability $p_i$. From the above mentioned methods, the "one-versus-one" method for multiclass classification problem has been employed in this work for its simplicity and ruggedness with the training dataset.

### D. Classifier Selection

The blood samples of this study contain highly overlapped cell cluster types, making linear discrimination between normal and abnormal samples not possible. However, with the use of kernels, this data can be transformed to high-dimensional spaces where linear separation can be easily achieved. Perfect candidates to deal with high-data overlap are Support Vector Machines (SVMs) and ANNs. Both are easy to model and have very simple training procedures. SVMs are a popular approach for two-category classification that is simple in use [2]. It is based on the structural risk minimization principle and focuses on finding the so-called support vectors [1] a set of critical points that can be used to represent the decision function of the classifier. They perform classification by transforming the data into a feature space by means of kernel functions. The drawback is that SVMs depend on a limited amount of kernels. As opposed to SVMs, ANNs can be extended to multi-category classification, but more importantly, they re-create feature spaces by simply changing neuron interconnections and activation functions, and therefore, each configuration represents a different kernel.

**Steps to work Procedure:**

1. In first step, we use image Acquisition
2. In second stage it will generate image preprocessing on tested data.
3. Then apply segmentation, feature extraction , and classification method
4. Identify the counting area ( the area where cells do not interfere with each other and are properly separated).
5. Perform color normalization on the image.
6. Separate white blood cells from the image by performing image segmentation.
7. Specify the region of interest.
8. Extract the region of interest.
9. Feature vector analysis is performed which includes extraction and selection of features of the cell images including the size of the cell, the circularity of cells, count of cells, the area of cytoplasm and nucleus, nucleus/cytoplasm ratio etc.
10. Initial classification of lymphoid or myeloid series is done by the shape of the nucleus.
11. The extracted features are then compared with those stored in the database and classification is done into AML or ALL categories.
12. K nearest neighbor, neural network and support vector machine classifier is used for the classification.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

The dataset used in this paper consisted of 100 actual microscopy images of blood samples. For AML, we accessed the *American Society of Hematology (ASH)* for their online image bank of leukemia cells.

 

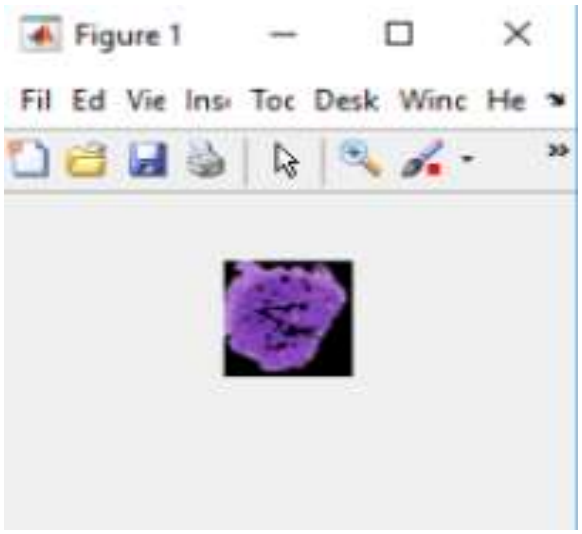


**Figure 2: Preprocessing phase of Leukemia**

**Figure 3: Hausdorff feature phase of Leukemia**

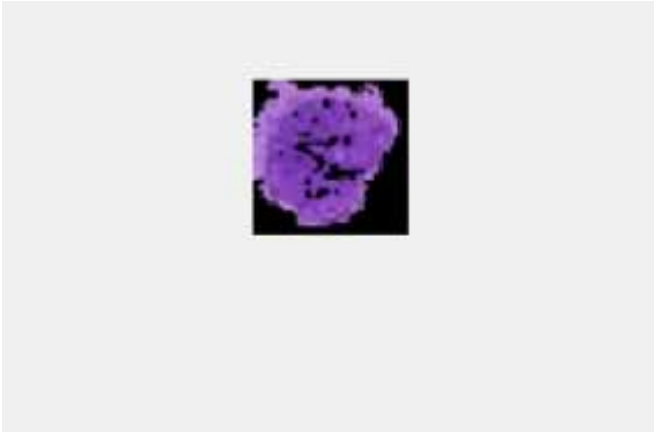

**Figure 4: LBP feature phase of Leukemia**

In LBP method each pixel is restored through a binary pattern that is derived from the neighborhood of the pixel.

 

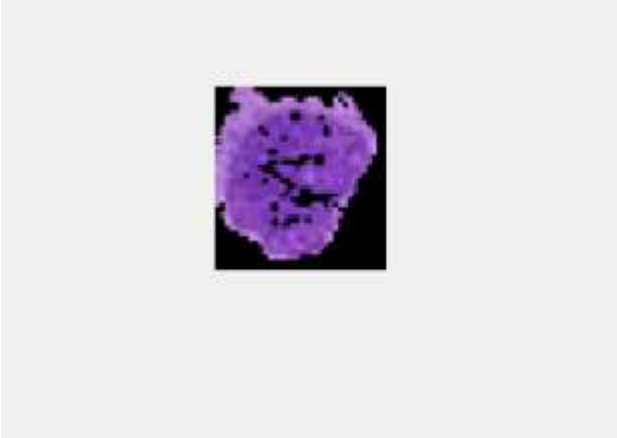

**Figure 5: Texture feature phase of Leukemia**

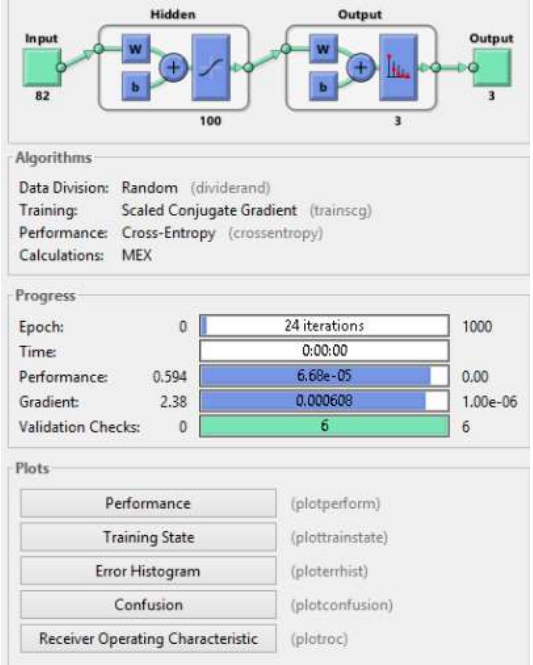


**Figure 6: NN training and testing phase**

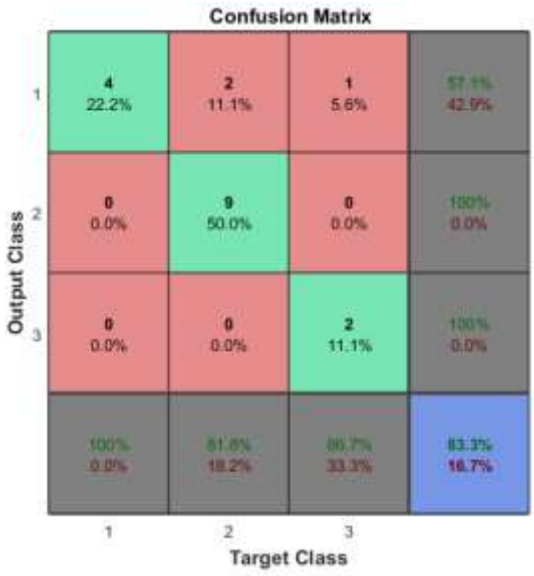


**Figure 7: Confusion Matrix**

 

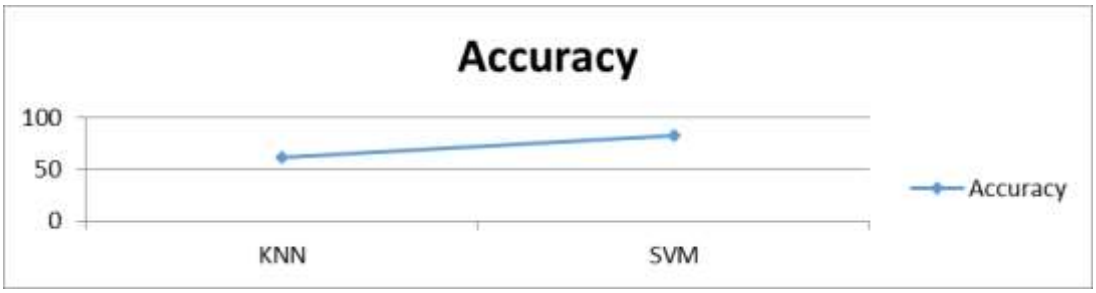


**Figure 8: Accuracy graph**

**Table 1: Accuracy Table**

| Method | Accuracy Percentage |
|---|---|
| **KNN** | **61.11%** |
| **SVM** | **83.33%** |

## CONCLUSION

This paper proposed a method to automate the segmentation, feature extraction and classification of red and white blood cells using KNN, NN, and SVM classification algorithm. Several improvements were made to the SVM algorithm, including an initialization step to find 12-neighbor connected component. Additionally, the proposed model features an enhanced accuracy of selecting the correct circle from three candidate circles, the capability to detect irregular cells, the use of a dynamic number of iterations, and improved detection of overlapping cells. The proposed method performed the segmentation and classification of WBCs and RBCs well when results were compared with the ground truth, which was determined by experts. The following segmentation and counting accuracies were achieved using the proposed method

## REFERENCES

[1] Cristianini, N., and J. Shawe-Taylor. "An Introduction to support vector machines and other kernel-based learning methods" New York: Cambridge University Press, 2000.

[2] Vapnik, V. N. "The Ature of Statistical Learning Theory" New York: Springer, 1995.

[3] A. Madabhushi, "Digital pathology image analysis: opportunities and challenges," *Imaging in Medicine*, vol. 1, no. 1, pp. 7–10, 2009.

[4] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 1, pp. 54–58, 2002.

[5] L. Yang, O. Tuzel, P. Meer, and D. J. Foran, "Automatic image analysis of histopathology specimens using concave vertex graph," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008*, pp. 833–841, Springer, Berlin, Germany, 2008.

[6] R. C. Gonzalez, *Digital Image Processing*, Pearson Education India, 2009.

[7] S. Liao, M. W. K. Law, and A. C. S. Chung, "Dominant local binary patterns for texture classification," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 1107–1118, 2009.

[8] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using a bag of features and kernel functions," in *Artificial Intelligence in Medicine*, vol. 5651 of *Lecture Notes in Computer Science*, pp. 126–135, Springer, Berlin, Germany, 2009.

[9] H. S. Wu, J. Barba, and J. Gil, "Iterative thresholding for segmentation of cells from noisy images" Journal of Microscopy, vol. 197, no. 3, pp. 296–304, 2000.

[10]C. Di Rubeto, A. Dempster, S. Khan, and B. Jarra, "Segmentation of blood images using morphological operators," in Proceedings of the 15th International Conference on Pattern Recognition, vol. 3, pp. 397–400, 2000.

[11]C.D., Ruberto, A., Dempster, S., Khan, B. Jarra, "Analysis of Infected Blood Cell Images Using Morphological Operators", Image and Vision Computing, Vol. 20, 2002, pp. 133-146

 

[12]Q. Liao and Y. Deng, "An accurate segmentation method for white blood cell images," in Proceedings of the IEEE International Symposium on Biomedical Imaging, pp. 245–258, 2002

[13] J.L. Starck, E.J. Candes, D.L. Donoho, "The curvelet transform for image denoising", IEEE Transaction on Image Processing 11 (6) (2002) 670–684.

[14] G. Lin, U. Adiga, K. Olson, J. F. Guzowski, C. A. Barnes, and B. Roysam, "A hybrid 3Dwatershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks," Cytometry Part A, vol. 56, no. 1, pp. 23–36, 2003.

[15] the Fabio Scotti University of Milan, Department of Information Technologies, via Bramante 65, 26013 "Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Microscope Images" IEEE International Conference on Computational Intelligence for Measurement Systems and Giardini Naxos, Italy, 20-22 July 2005.

[16] Subrajeet Mohapatra, Dipti Patra, Sanghamitra Satpathy,"Automated Leukemia Detection in Blood Microscopic Images using Statistical Texture Analysis",inICCCS, February 2011.

 