Proceedings of the 2nd International Conference on Inventive Computation Technologies (ICICT 2017)
IEEE Xplore Compliant - Part Number:CFP17K52-ART, ISBN:978-1-5090-6697-1

Implementation of Computer Aided Leukemia Detection using Digital Image Processing Techniques

Shaikh Mohammed Bilal N, Department of Computer Engineering, University Of Mumbai, Bilals.sb@gmail.com
Sachin Deshpande, Department Of Computer Engineering, University Of Mumbai, Sachin.deshpande@vit.edu.in

***Abstract*—The machine-controlled identification of malignant cells from microscopic diagnostic assay pictures of blood samples helps in assuaging the diagnostic problems of leukemia and provides higher results if the biologically explainable and clinically important feature primarily based approaches are used for the identification of malignancy and the severity of the disease. Identification method might have error rates up to 40%; Due to this drawback it is necessary to develop a second opinion tool which helps the doctors and pathologists in a classified detection of the leukemia disease. The analysis shows a discourse method for leukemia and its subtypes detection from microscopic blood cell images which can be done by applying image processing techniques followed by a detailed classification analysis on the dataset images.**

Keywords—leukemia, detection, segmentation, feature extraction, classification, subtypes.

I. INTRODUCTION

Leukemia is a type of cancer that originates in the bone marrow of the subject. It occurs as a result of large production of immature leucocytes which override the normal blood cells (WBC, RBC, and platelets). The body gets exposed to multiple other diseases due to the decreased WBC efficiency to fight against the same. within the diagnosing of cancer, additionally to think about the signs and symptoms of the subjected patient, it's needed to identify the malignant or blast cells. To get the amount of the blood cells in every unit volume a blood count is calculated for different types of cells in blood(RBC, WBC, and platelets). For the determination of the number of malfunctioning leucocytes, a bone marrow cellular analysis is carried out after looking for abnormalities in the blood count. A diagnostician examines the microscopic cell images in correlation to a research to determine the abnormalities in the leucocytes in order to observe the occurrence of blood cancer and further its type and subtype. Classification decides the prescription for the subjected patient of blood cancer, hence it is very important. Although Flow Cytometry is used by various pathologists as a dependable mechanism for blood cancer diagnosis, still in many medical institutes an hospitals it is not feasible, mainly the public sector medical hospitals, as described in [10]. During this work, the type and subtype of Acute Leukemia can be identified by working upon the information provided by the microscopic blood cell images of the dataset of different human subjects or patients.

II. PROBLEM DEFINITION

A. Statement Of Problem

To analyze blood cells digital pictures from blood smear samples, then perform a discourse approach for the detection and classification of Acute leukemia subtype after determining the type of Acute leukemia.

Acute leukemia is of two types namely, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloblastic Leukemia (AML). Further, Acute Lymphoblastic Leukemia (is divided into the subsequent subtypes:

1. ALL subtypes are L1, L2, and L3.

B. Exiting Detection Systems

Because of the irregular staining and overpopulated cell smears the segmentation of cytoplasm and nucleus is difficult. SVM along with Discrete Fourier Transform and segmentation was used in some works, which in turn provided an automated learning mechanism to differentiate the blood smear image specifications (i.e cytoplasm cells, nucleus and platelets). This method is more acceptable and efficient in contrast with the "Thresholding and the Watershed Algorithms", as in [1]. Moreover, a similar version of this method named as "Simulated visual attention via learning by on-line sampling" is projected in [2]. Due to the issue of overlapping cells in blood images some algorithms were introduced. They usually worked at splitting and merging cells at the edges, as shown in [3]. Detailed classification was deployed by Mohapatra et. al. using various classifiers to classify the healthy and cancerous cells, which is stated in [4]. Lim et. al. projected on a method thereby showcasing "thresholding, morphological operations, and watershed mechanism" as shown in [5].

III. LITERATURE SURVEY

Table 1

Sr. No.	Paper	Author	Advantages	Drawbacks
1.	**"Computer Aided Detection of Skin Cancer" [7][15]**	Aswin R. B. and Abdul J.	Hybrid Classification Approach, Effective feature extraction [7].	Cell overlapping not addressed, Can be used for detecting melanoma only, Accuracy is comparatively less
2.	**"Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features" [6] [15]**	Rajesh K.	Texture, shape and morphology, HOG, wavelet color, Tamura's feature, and LTE used innovatively with an accuracy of 92 % [6].	It was observed that the proposed method is performing better for connective tissues type sample and was not tested for leukemia blood smear samples.
3.	**"Detection Of Skin Cancer Using Hybrid of SVM-ID3 Algorithm" [11]**	Greeshma Rajan1, G Shivaraj	Cell segmentation, Third Harmonic Generated Microscopy, SVM, ID3, Hybrid of ID3-SVM Algorithm, Nucleus-to-Cytoplasm (NC) ratio implemented in a simple effective manner [11].	Cell overlapping not addressed, Texture features not addressed.
4.	**"Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms" [12] [15]**	Ammara M.	Detailed comparison and classification of various models used in diagnosis of cancer such as ultrasound and optic coherence tomography [12].	Limited morphologic Information. Blood cancer not involved.

As shown in Table 1, Aswin. R.B and J. Abdul et al. in "**Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features"** 2014(IEEE) used a hybrid GA ANN classifier after applying image pre-processing (dull razor & filtering), segmentation (color threshold seg) and feature extraction (grey level co occurrence matrix-GLCM) techniques on images of skin for melanoma detection with an accuracy of 88% [7].

Rajesh Kumar in his paper "**Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features**" 2015 showed an approach to detect and classify cancer from blood cell images specific pathological features, which is shown in [6].

For segmentation of images color *k*-means based method is used. The various hybrid features which are extracted from the segmented images include shape and morphological features, GLCM texture features, Tamura features, Law's Texture Energy based features, histogram of oriented gradients, wavelet features, and color features. For classification purposes, *k*-nearest neighbor based method was proposed to be used with an average accuracy of 92.19%.

IV. IMPLEMENTATION

Implementation was completed using MATLAB r2014a which incorporated image processing algorithms followed by classification using SVM (Support Vector Machine). The training of the dataset is carried out using ANN (Artificial Neural Network).

A. Algorithm 1: Image processing

This includes procedures for Feature Extraction, Filtering and Clustering.

1. Feature Extraction:

Feature Extraction is carried out to extract certain specific features from the image of the dataset [3]. It includes the following operations:

a. Image Resizing:

The input image is resized to a 256 x 256 resolution image using
I = imresize(I,[256 256]); function.

b. HSV image:

The resized image is in RGB standard format which is converted to HSV format using the function *H = rgb2hsv(I);*
Hue values are parameterized to be between intensity values 0.6 to 0.9 and saturation values are adjusted to be <0.3 as follows:

```
mask=(H(:,:,1)>0.6);
mask1 =(H(:,:,1)<0.9);
mask2= H(:,:,2)>0.3;
```

Where mask1 is in the hue plane and mask2 is in the saturation plane, later on both the masks are combined to obtain a clear HSV image using:

```
mask = mask.*mask1;
mask=mask.*mask2;
```

c. Median Filtering:

Filtering is usually done to remove the noisy data from the images for more accurate processing results [10]. The extraction of nucleus and cytoplasm plays an important role in the diagnosis of the presence blast cells. The grayscale of the nucleus is obtained using:

```
I = rgb2gray(I);
Median filter procedure:
  %MF of the image I
    Aci=trapmf_mat(I,a1,a2,a3,a4);
    TM(:,:,iNmax)=Aci;
end
   [Mx My Mz]=size(TM);
 [Vx S]=max(TM,[],3);%AGREGRATION by MAX
 if MODE==0            %Max Median aggregation
    if (exist('argum','var'))
       Ws=argum;
    else
       Ws=3;
    end
    MG=medfilt2(S,[Ws,Ws]);
```

Followed by calling:

```
I = medfilt2(I);
```

d. Fuzzy C-means Clustering:

The segmentation of the nucleus and cytoplasm as well as to determine the overlapped morphologies Fuzzy C-means Clustering is used based on 8 Neighborhood pixel matrix for cluster formations of the image. 90% of the malignant data is composed in the nucleus of the blood stains hence Fuzzy C-means proves the fast and efficient procedure for clustering of the same [3].
The procedure is as follows:

```
function[MG,S,TM,Nmax,Nth]=fuzzycmeans(I,Nth,method,sm
ooth,OPTS)

if(~exist('Nth','var'))
   Nth = 0;
end
if(~exist('method','var'))
   MODE = 1;
   Niter=  3;
else
   if length(method)==2
      argum=method(2);
      MODE=method(1);
   else
       MODE=method;
   end
end
if(~exist('smooth','var'))
   smooth=1.5;
end
if(~exist('OPTS','var'))
      TH=0.008;
else
  TH=OPTS;
end

I=double(I);
Followed by calling:
[fcimg,Nth]= fuzzycmeans(I,0,1);
```

Thereafter, applying a binary mask on FCM output image we get Binary FCM Image as:

```
fcmsk = (fcimg ==max(max(Nth)));
```

e. Texture Analysis:

Thereafter Gabor filtering takes place at 4 x 4 pixel matrix level. This is done in order to determine the texture of the segmented nucleus and cytoplasm. These filters are similar to the Human Visual Perception used for texture representation [9].

```
gaborArray = gaborFilterBank(5,8,39,39);  % Generates the
Gabor filter bank
```

GaborFilterBank is the decomposition of the image (generated from the previous step) into 5 scales and 8 orientations and the rows and columns are set to be 39 of the 2D Gabor Filter [10]. This *GaborArray* is used to extract the Gabor features to generate a feature vector using:
featureVector = gaborFeatures(img,gaborArray,d1,d2)
fet= max(featureVector)

Finally the maximum of the feature vector is obtained which consists of specific features that will be used to target the class of SVM Classification process as:

FEAT = [std ; stdm; arm; armin ; fet;Centroid ;Orient; ... Extent;Solidity;Perimeter;Eqdiameter;Convexarea;Majaxislength; Minxislength];

The list of the feature vectors is as follows:

1. *areag(j1)= nnz(n1);* //**Area of image**
2. *stddev= sqrt((1/N)*sigm*); //**Standard deviation**
3. *reg = regionprops(n1,'all');* //**Region of interest**
4. *Centroid= reg.Centroid(1);* //**Centroid of ROI**
5. *Orient= reg.Orientation;* //**Orientation ROI**
6. *Extent= reg.Extent;* //**ROI extent boundary**
7. *Solidity= reg.Solidity;* //**Solidity of ROI**
8. *Perimeter= reg.Perimeter;* //**Perimeter of ROI**
9. *Eqdiameter= reg.EquivDiameter;* //**Diameter of nucleus**
10. *Convexarea= reg.ConvexArea;* //**Convex area of ROI**
11. *Majaxislength = reg.MajorAxisLength*; //**Major axis length of ROI**
12. *Minxislength= reg.MinorAxisLength*; //**Min axis length of ROI**
13. *std= max((stddev));* //**Max std deviation**
14. *stdm= min(stddev);* //**Min of std deviation**
15. *arm= max(areag);* //**Max Area**
16. *armin= min(areag);* //**Min Area**

B. Algorithm 2: Training and Classification:

1. Training:

Training of the dataset is done using Artificial Neural Network (ANN). Automatic Maxima Search is used for training which determines the number of clusters formed in each image based on features extracted by each image of the dataset.

```
if Nth==0 %Automatic maxima search
   Nth=maxima_search(I,TH);
   disp(['Number of clusters: ' num2str(Nth)])
```

Training is followed by the formation of graphs which visualize the the clustering of the dataset classes using Principal Component Analysis (PCA).
PCA is a process used to determine the variations in patterns of a dataset and thereafter visualizing it in a graphical manner for the ease of analysis. **Principal component analysis** (**PCA**) is a graphical process which uses an orthogonal conversion to transform a set of observation values of possibly related variables into a set of values of linearly unrelated variables called **principal components** [6].

```
%PCA analysis
[P,score]=pca(FEAT,'NumComponents',2);
```

Where P is p by p matrix, each column containing coefficients for one principal Component and score is the data formed by transforming the given data into the space of principal components.
PCA is followed by K-means Clustering to obtain the centroids and indices based on the various principal components obtained using PCA. This ultimately results in a graph containing four groups or clusters as “NO leukemia”, “L1”, “L2”, and “L3”; as shown in the fig 1.

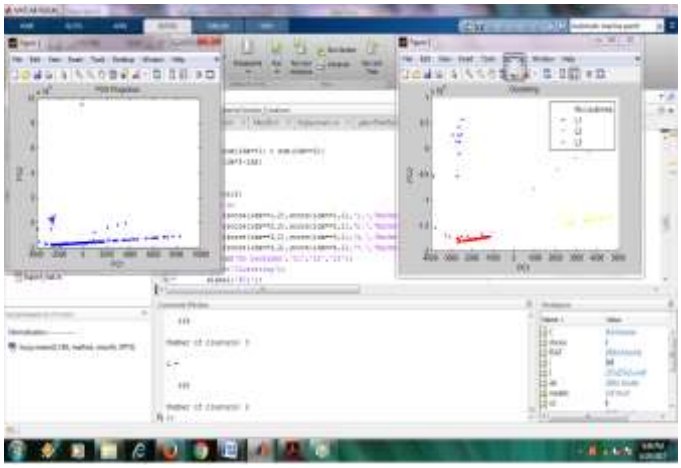

Fig 1. Clustering and PCA graphs obtained by training of the Dataset.

2. Classification:

Testing of the trained dataset is achieved using Support Vector Machine (SVM). The models for classification are generated using *svmtrain* function along with the number of clusters obtained after PCA.

```
%SVM train
   [models,nC] = multisvmTrain(FEAT,idx);
   save('data.mat','models','nC');
```

These models are saved and are used for classification along derived from the feature vector and indices obtained in PCA. We use *multisvmTrain* function of MATLAB since our approach focuses on the diagnosis as well as the subtype diagnosis of Acute Leukemia.

Support Vector Machine (SVM)

In machine learning, **support vector machines** are supervised learning methods that perform data analysis for classification. Input is a set of training samples of the dataset, each belonging to either of the two class, an SVM training algorithm makes a model to assign new samples to one class or the other, as a non-probabilistic binary classifier [7].

Multiclass SVM

We use Multiclass SVM algorithm for the classified diagnosis of our dataset. The objective behind using such an approach is it aims to assign class labels to instances which are derived from several elements of the dataset. This approach is useful in eliminating the single multiclass problem by multiple binary classification technique. This results in building binary classifiers which distinguish between one and the rest of the labels -One v/s All. Classification in the course matter of One v/s All method is done on the basis of 'Winner-takes-all-strategy, in which the classifier with the highest output function assigns the class, comparatively. It involves training a single classifier per class, assuming samples of that class as positive and all other samples as negatives.

result= multisvmClassify(models,numClasses,TestSet)

The One v/s all strategy is implemented as:

```
%classify test cases
for j=1:size(TestSet,1)
   for k=1:numClasses
      if(svmclassify(models(k),TestSet(j,:)))
         break;
      end
   end
   result(j) = k;
end
```

The following figures show the Outputs of the detection:

Fig 2 Output showing ALL detected with subtype L1.

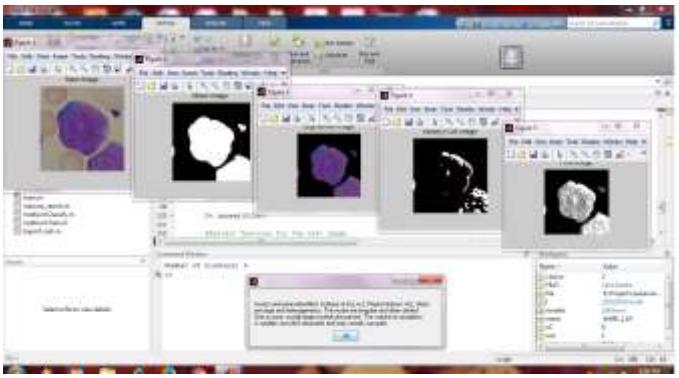

Fig 3 Output showing ALL detected with subtype L2.

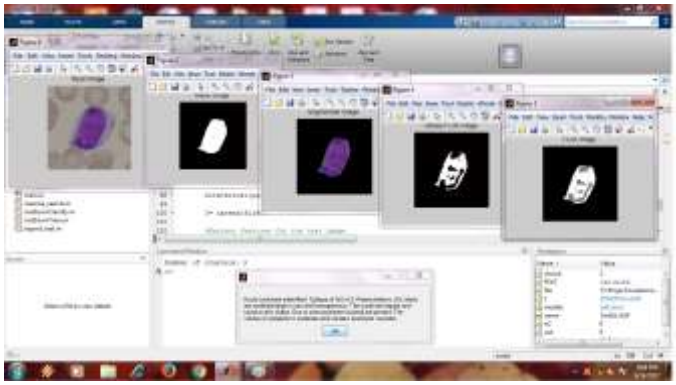

Fig 4. Output showing ALL detected with subtype L3.

Fig 5. Output showing NO leukemia.

V. RESULTS

The criteria for evaluating the classified diagnosis of selected 63 images with high visual variations from the dataset were: overall percentage of correct classifications, true positive rate (TPR) and true negative rate (TNR). The metrics used for evaluating the classification were: Precision = TP / P, FP Rate = FP / P, and FN Rate = FN / N, where TP and FP are the number of images correctly and incorrectly classified as leukemia respectively, P is the number of images classified as having leukemia, FN is the number of images incorrectly classified as NO leukemia, and N is the number of images classified as NO leukemia.

With respect to the diagnosis algorithm, as shown in table 2; we found that by using multiclass SVM classifiers it was possible to reduce the number of false positives and false negatives that were present. This allowed us achieving an accuracy of 95% in the diagnosis of acute leukemia families and its types. Evaluation of the acute leukemia diagnosis algorithm by multi-class SVM classifier results is shown in the table below:

Classification	No. of images	Correct% (TP&TN)	Failed% (FN&FP)
Types	63	95.22	4.78
ALL	43	90.69	9.31
NO Leukemia	20	99.75	0.25
Subtypes			
L1	15	86.66	13.33
L2	14	85.71	14.29
L3	14	100	0

Table 2 Classification Results

VI. FIELD WORK

The **TATA Memorial Hospitals (TMH**), Mumbai was visited by me as part of the project detailing and research and analysis of the topic.

It was reported by the senior pathologist of the **Hematology** department that for the blood cells count and determination of abnormalities they used **flow cytometry** method implemented by Beckman Coulter imported by- Beckman Coulter, Inc.

The **COULTER VCS** uses unit volume, conductivity and scatter light parameters for the determination of WBC and its diagnosis, which is shown in the manual at [9], and elaborated as follows:

Volume Analysis: The unit volume of the individual blood cells is evaluated by the Beckman Coulter using very low frequency current. This method has been used since ages to evaluate the WBC count in a blood sample, which is further processed for the diagnosis of leukemia.

Conductivity Analysis: High frequency current passes easily through cell walls of the blood. As they are good conductors, the evaluation of insulation at the walls of the cell as well as the specifications of the nucleus components along with the chemical uptake and traces inside the cell is evaluated.

Light Scatter Analysis: Scattering of laser light throughout the population of cells contained in a blood sample of a patient also known as Fulwyler's method for cell analysis, is deployed in Beckman Coulter so as to determine the blast cell count, the containment proportion of the cells and the overpopulated regions.

VII. DATASET

As urged by the doctors of the Tata Memorial Hospital, Mumbai; the request to accumulate the dataset was sent to Fabio Scotti, University of Milano, Italy. The database ALL IDB was gathered that consisted of ALL_IDB1 and ALL_IDB2 datasets consisting of pictures that were captured by optical laboratory magnifier as well as a Canon PowerShot G5 camera. The resolution of the JPEG format 24 bit pictures is 2592 x 194, which is stated in [13]. ALL_IDB1 consists of 108 pictures (healthy and malignant) having 39000 blood components in which the lymphocytes are labeled by skilled oncologists.ALL_IDB2 dataset is for testing and training the classification system. It's a set of separated regions to be considered pertaining to healthy and leukemia cells from ALL_IDB1 dataset.

Acknowledgment

I wish to offer my sincere gratitude to our principal Dr. S.A. Patekar for providing me an opportunity to be a part of the Vidyalankar Institute of Technology, Mumbai under the postgraduate program and moreover to allow me work on this final year Project on the respective domain.

I sincerely thank my Guide Prof. Sachin Deshpande for the guidance and encouragement in carrying out this work.

I would like to offer gratitude to the pathologists of the Hematology Department, TMH, Mumbai for their support and information gathering.

References

[1] Pan C, Lu H, Cao F. "Segmentation of blood and bone marrow cell images via learning by sampling." Emerging Intelligent Computing Technology and Applications, Springer Berlin Heidelberg. 2009; 336–345.

[2] Pan C, Park DS, Yoon S, Yang JC. "Leukocyte image segmentation using simulated visual attention." Expert Systems with Applications. 2012; 39(8): 7479–7494.

[3] Wang W, Hao S. Cell cluster image segmentation on form analysis. IEEE Third International Conference on Natural Computation. 2007; 4: 833–836.

[4] Mohapatra S, Patra D. "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images." Neural Computing and Applications. 2014; 24(7–8): 1887–1904.

[5] Lim HN, Mashor MY, Hassan R. "White blood cell segmentation for acute leukemia bone marrow images." IEEE International Conference on Biomedical Engineering (ICoBE) 2012; 357–361.

[6] Aswin.R.B and J. Abdul "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features." 2014(IEEE).

[7] Rajesh Kumar," 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, 2015.

[8] Weka—Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update; ACM SIGKDD Explorations Newsletter. 2009; 11(1): 10–18.

[9] COULTER HmX Hematology Analyzer, PN 4237523CB (March 2011) Beckman Coulter, Inc.250 S. Kraemer Blvd.Brea, CA 9282.

[10] Paschos G. Perceptually uniform color spaces for color texture. IEEE Transactions on Image Processing. 2011; 10(6): 932–937.

[11] Greeshma Rajan et al, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, May- 2014, pg. 143-153.

[12] Ammara Masood, "Computer Aided Diagnostic Support System for Skin Cancer:A Review of Techniques and Algorithms" Hindawi Publishing Corporation, International Journal of Biomedical Imaging Volume 2013, Article ID 323268.

[13] F. Scotti "ALL_IDB: the acute lymphoblastic leukemia image database for image processing", 2011 IEEE, Brussels, Belgium, pp. 2045-2048, ISBN: 978-1-4577-1302-6.

[14] www.semanticscholar.org / http://dx.doi.org/10.1155/2013/323268

[15] C. Reta. " Leukocytes Segmenta ion Using Markov Random Fields" , Advances in Experimental Medicine and Biology, 2011.