Journal of Theoretical and Applied Information Technology
30th September 2020. Vol.98. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STATISTICAL MORPHOLOGICAL ANALYSIS BASED SUPERVISED CLASSIFICATION ALGORITHM FOR DIAGNOSING ACUTE LYMPHOBLASTIC LEUKEMIA

[1]JKC SHYALIKA, [2]PPNV KUMARA, [3]DU KOTTAHACHCHI

[1]Graduate, Department of Information Technology, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka

[2]Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka

[3]Department of Medical Laboratory Sciences, Faculty of Allied Health Sciences, General Sir John Kotelawala Defence University, Sri Lanka

E-mail: [1]chathurangijks@gmail.com , [2]nandana@kdu.ac.lk , [3]darsha.uda@gmail.com

ABSTRACT

Leukemia is a fatal disease of the type "Blood Cancer", where the White Blood Cells (WBC) increases in human bone marrow and peripheral blood. Acute Lymphoblastic Leukemia (ALL) is a common types of leukemia that affects young children of below 10 years and adults over 60 years, aroused by accumulation and overproduction of immature and cancerous cells identified as lymphoblasts. At present, the diagnosis of ALL includes measures alike performing a full blood count, bone marrow biopsy, blood picture, immunophenotyping, cytochemical stain and cytogenetics. These medicinal techniques are highly tedious, costly, requires expertise of hematologists and available only in few hospitals especially in developing countries. Hence, as an alternative, use of image processing and machine learning to diagnose ALL would become an effective solution. Even though, several research groups have used image processing to detect and diagnose ALL, recognition and splitting of overlapping Red Blood Cells (RBC) with WBC has however been a challenging issue. This paper is about a research study and an application that includes an image processing and machine learning algorithm to diagnose ALL while attempting to solve the issue of overlapping cells. The research is also extended to detect the quality devastation in blood films in terms of storing them for prolonged period. The inputs for this application include microscopic peripheral blood films of ALL patients and healthy individuals obtained from Department of Pathology Clinic at Faculty of Medicine, University of Colombo, Sri Lanka. This research project has received verification of ethical approval from Faculty of Medicine, General Sir John Kotelawala Defence University, Sri Lanka. In the developed application, segmentation using morphological operations in OpenCV Python and supervised learning based classification using K-Nearest Neighbour implementation has been proposed in detection and diagnosing of ALL. As per the results, the proposed algorithm has led to a high accuracy of 88.8% in diagnosing ALL. The end product includes a Python based QT GUI based development suite that performs main targeted backend functionalities and a PHP based web application that serves hematologists, doctors and patients to perform utility functions.

Keywords: *Acute Lymphoblastic Leukemia, Image Processing, Segmentation and Feature Extraction, Classification, K-Nearest Neighbour, Supervised Learning*

1. INTRODUCTION

Blood is one of the most important materials of the human body as it is the prime agent that make humans live. Human blood consists of two major parts; plasma and cells. Plasma consists of 90% water and other compounds such carbohydrates, proteins, hormones, lipids, electrolytes [1, 2]. In adults, cells originate from the bone marrow in a specialized cell type known as stem cells and transferred to periphery when they become matured. The peripheral blood cells consist of three main components; Red blood cells (RBC), White blood cells (WBC) and platelets. There are five types of WBCs; Neutrophils (40-75%), Lymphocytes (20-45%), Eosinophils (1-8%),

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Monocytes (0-10%), Basophils (0.5-1%), Neutrophils, Eosinophils and Basophils have granules in their cytoplasm; hence they are named as granulocytes. Monocytes and Lymphocytes do not have granules, thus are termed agranulocytes [1].

Leukemia, simple called "Blood cancer" in which usually the number of WBC increase in the bone marrow and peripheral blood. These leukemic cells (usually immature) replace the other normal blood cells causes malfunction of the bone marrow as well as peripheral blood. Furthermore, excess amount of these cells travels to other sites such as liver, spleen to maintain normal cells production. Later, the leukemia cells also invade other organs causing them to malfunction [1].

There are two main types of leukemia according to the morphology of cells in the bone marrow. They termed as acute & chronic Leukemia. Generally, acute leukemia involves the rapid overgrowth of very immature blood cells whereas chronic leukemia involves the overgrowth of somewhat mature blood cells in the bone marrow compared to acute type. With the introduction of French-American-British (FAB) classification in 1976 [2], acute leukemia further categorized into two groups based on the white blood cell from which the malignancy originates from. They are Acute Lymphoblastic Leukemia (ALL) is caused by abnormal lymphoid cells, and Acute Myeloid Leukemia (AML) is caused by abnormal myeloid cells in the bone marrow [1]. The predominant abnormal cells in the ALL are lymphoblasts.

Diagnosing leukemia usually begins with a medical history and physical investigation. If leukemia is suspected, the patient is made to undergo a number of tests in order to detect and diagnose leukemia and also to identify the sub type. These tests include, performing Full Blood Count (FBC), Blood Picture (BP) to identify abnormalities in cell shape, a bone marrow aspiration and trephine biopsy is then conducted to identify the type of abnormal blood cells [3], cytochemical stains to demonstrate enzymatic activity, carbohydrates or lipids present or any other special characters present in leukemia cells [2]. To clarify the subtypes and also for comprehensive diagnosis, an advanced technique such as immunophenotyping and cytogenetics are employed [4, 5]. The whole process takes about 3-4 days and also needs well-trained experienced professionals to supervise. However, early diagnosis of leukemia contributes to early treatment and proper management of patients. Furthermore, manual detection procedure stated above is a highly tedious task that involves the effort of hematologists and other supporting staff as it is intensively slow, costly, time consuming. Even though advanced techniques are being used, there may be errors especially diagnosing subtypes.

Image Processing and Machine Learning fields have provided fast, cost effective and accurate solutions in fields such as medical image management, image data mining, bio imaging, neuroimaging and virtual reality in medical visualization [6, 7]. Image processing techniques includes Image acquisition, restoration, pre-processing, segmentation, Feature extraction, compression, wavelets, representation, recognition etc. [7]. A digital image is a representation of a two-dimensional image as a finite set of digital values called pixels. Image processing is a type of signal processing method that perform some operations on these digital images in order to get an enhanced image or to extract some useful information from it [4]. Machine learning is the branch of Artificial Intelligence that provides systems the ability to automatically learn and improve from experience that is without being programmed explicitly. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. However, there are some challenging issues in these areas. There are unsolved problematic areas such as quality degradations occurring in image compression and enhancements, in recognizing generic objects, visualization issues etc. [6, 7].

Researches have been conducted for the detection and counting of RBC [8], white blood cells and to diagnose diseases like anaemia, malaria and deficiency of vitamin B12 using blood images [9]. Furthermore, Image Processing techniques have been used for detecting cancer cells [10]. For Image Classification, Supervised or Unsupervised Machine learning techniques have been used [8, 10, 11]. The objective of this paper is to present the results of a research project conducted in order to diagnose a type of blood cancer, (ALL). In this research project, an efficient and accurate image processing algorithm to detect and diagnose ALL cells using microscopic images obtained from human blood peripheral blood films stained with Leishman's has been proposed. The proposed system would use Digital Image Processing and Machine learning techniques in order to complete the task.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Section 2 of the paper discusses the early research attempts on diagnosing ALL that ultimately reveals the research gap for ALL diagnosis. Section 3 is followed by the methodology and the experimental design of the solution developed by the authors. Section 4 presents the evaluation results incurred as the research outcomes of the system testing. Section 5 concludes by indicating the research achievements and identified research challenges for ALL diagnosis.

2. EVOLUTION AND STATE-OF-THE ART IN ALL DIAGNOSIS

Presently, a considerable contribution has been done by researchers in the aim of ALL diagnosis using image processing and machine learning. The common flow of the image processing techniques that is used in diagnosis can be illustrated by Figure 1.

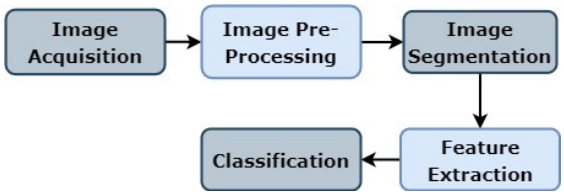


Figure 1: Flow of ALL Diagnosis

The researches done so far varies from the segmentation methods and classification methods they have used. A comparative review on the segmentation methods and classification methods deployed by early researchers in diagnosing ALL have been elaborated in this section.

2.1. Segmentation Methods

Segmentation process partitions an image to its constituent segments or objects known as pixels. This locates objects and boundaries (curves and lines etc.) of images and modifies the representation of an image into somewhat that is more meaningful and easier to analyze. In literature, segmentation has been used to separate the WBC from the cytoplasm, identify the leucocytes and their nuclei, identify grouped leucocytes and for image cleaning. Mostly used segmentation methods used in leukemia detection have been discussed in this section.

2.1.1. Watershed segmentation

"Watershed" refers to a ridge that splits areas drained by different river systems [11]. Watershed lines are defined on the nodes, edges, hybrid lines on both nodes and edges and in continuous domain. Watershed segmentation is an easy method for the detection of WBC but requires best quality images in order to achieve a better accuracy [11].

2.1.2. Fuzzy C Means clustering (FCM)

This data clustering technique groups a dataset into n clusters with all data points in that dataset belong to each and every cluster to a certain degree. FCM result is much accurate and it's able to measuring nucleus boundaries with shape, colour and texture, but it's difficult in classification of lymphoblast in to its sub types through this segmentation [11,12].

2.1.3. Fuzzy K-Means clustering in L*A*B* colour space

K-Means method is a least squares partitioning method and it divides a collection of objects to K groups of clusters. It considers each object have a location in the space and finds partitions in the image such that objects within each cluster close to each other as likely, and as far from the objects in other clusters as possible. This method is not applicable on incremental data and it cannot give classification with labelled data [9]. Mohapatra and the colleagues have used fuzzy based blood image segmentation for separate out leucocytes from other blood components [13].

2.1.4. Otsu's method

This is a thresholding method and it's the easiest and fastest method used in segmentation. Thresholding is based on a clip-level named a threshold value used in converting a grayscale image into a binary image. Fabio Scotti has used Otsu's method in nucleus and cytoplasm selection in lymphoblasts and lymphocytes [14]. Their experiments have showed a good performance of this method in separating the nucleus from the cytoplasm.

2.1.5. Shadowed C-Means clustering (SCM)

SCM is a method of partitive clustering developed in the framework of shadowed sets. Unlike rough clustering, in SCM, the choice of threshold parameter is fully automated and the number of clusters is optimized in terms of various validity indices [15]. Shadowed clustering can handle overlapping among clusters efficiently and also it

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

can model uncertainty in class boundaries [16]. The algorithm is robust in the presence of outliers too. However, fuzzy c-means clustering have problems with high dimensional data sets and a large number of prototypes [17].

2.1.6. HSI colour based segmentation

HSI (Hue, Saturation and Intensity) is a common colour model used in image segmentation. HSI colour model has a good capability of representing the colours of human perception [18]. Nor Hazlyna and the team have conducted a research for ALL detection based on segmentation using HSI and RGB colour space [19]. The results have shown that the proposed segmentation technique based on HSI has successfully segmented the acute leukemia images while preserving significant features and removing background noise. Singhal and Singh [20] and Halim and his colleagues [21] are some research groups who have used HSI colour based segmentation in ALL diagnosis. They have used HSI colour based segmentation as it provides better performance than RGB colour segmentation.

2.1.7. K-Means clustering

K-means clustering is an unsupervised learning algorithm which involves two simple processes as relegating the given data set and classifying the colligated data sets to the centroid nearest to them. K-means clustering segmentation have been used in identifying the leukemia sub types [22,23] and in AML screening systems [24]. K-means clustering does not give classification with labelled data and also not applicable on incremental data [9].

2.1.8. Morphological operations (shape-based)

Segmentation using morphological operations is a technique considering the processing of geometrical structure based on set hypothesis, topology, lattice hypothesis and arbitrary functions etc. This is the most successful segmentation method that has been used so far. Through this method it is very easy for detecting white cells, overlapping of cells and shape of cells. Thus this is based on statistics so can get approximate results.

Bhattacharjee and Saini [25], Vaghela et al.[9] and Raje and Rangole [26], Shyalika, Kumara and Kottahachchi [27] are among the researchers who have used morphological based image segmentation in leukemia diagnosis. They have discovered that the morphological operators used for the extraction of features have resulted in high segmentation accuracy. Segmentation using morphological operations has been used in morphological classification of Leucocytes by microscopic images [13,28]. In these researches, the researchers have focused on reducing the problem of identification and classification of WBC types in microscope images using morphological operations.

Mostly used segmentation techniques in ALL diagnosis and the advantage/merits and disadvantages/demerits of the varied techniques have been highlighted in Table 1. As per the evaluation results in earlier researches, Morphological/shape-based segmentation can been verified as the best method for image segmentation in ALL detection.

Table 1: Merits and Demerits of Segmentation Techniques.

Method	Merits	Demerits
Watershed transform	Easy method for detection of white cells	It cannot give accurate result and cannot implement on each and every image
K-means clustering	It is used for clustering and separate the data based on value of K.	It does not give classification with labeled data and also not applicable on incremental data.
Edge detection using histogram equalizing method and linear contrast stretching	This is very useful method to detect white cell and for contrast enhancement.	It is hard to define boundary of overlapping cell.
Shadowed C-means clustering(SCM)	SCM can handle overlapping among clusters efficiently and can model uncertainty in Class boundaries. Robust in the presence of outliers.	Have problems with high dimensional data sets and a large number of prototypes.
Shape based features	Very easy for the detection of white and overlapping cell and shape of cell	This is based on statistics so can get approximate result.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2. Classification Methods

Classification in Machine Learning and Statistics is a supervised learning approach in which the program learns from the input data and then uses this learning to classify new observations. It is in charge of assigning to the unknown test vector for new observations which is a label from one of the known classes [29]. Mostly used classifiers are discussed in this section.

2.2.1. Support vector machine (SVM)

SVM is a discriminative classifier that is formally defined by a separating hyper plane. When labelled training data is given (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. Dodandeniya, Kumara and Kulasekara in their research on Automated Blood Counter [8] have used SVM to separate the white cells and form red cells. Patel and Mishra [30] is a research group who presented an automatic approach for leukemia detection using microscopic images. Colour, geometric, shape and statistical features have been analyzed and classified under the SVM classifier in the intention of grouping the normal and abnormal cells. SVM has been used to classify leukemia types too. A three-layered framework consists of feature extraction, coding, and classification for the detection of leukemia from blood smear images has been proposed by Faivdullah and his colleagues [31] leukemia types. They have employed a one-vs-all technique to convert SVM, which is a binary classifier in to a multi-class classifier.

2.2.2. Artificial neural network (ANN)

This is a statistical learning algorithm defined by an interconnected set of nodes that are similar to the network of neurons found in brain. ANNs are capable of pattern recognition and machine learning, thus is mainly used in generating and estimating the output from a large number of input data set [25]. Mohapatra and the colleagues [32] have engaged in another project in Lymphocyte image segmentation using Functional Link Neural Architecture for ALL detection [32]. Fatma and Sharma [22] have tried on a system to identify and classify sub types of acute leukemia using neural network.

2.2.3. CART (classification and regression trees)

CART (Classification and Regression Trees) statistical method has been used in automatic leukemia diagnosis in investigating the classification power of cell markers extracted in segmentation [33]. This method generates classification tree diagrams with complete splitting information at each node and then produces a classification matrix, splitting cost and probability matrix for both the learning sample and the cross validation. The classification trees can be saved and used in classifying unknown specimens. Serbouti and the research team has employed CART in their research done in automatic leukemia diagnosis [33].

2.2.4. K-Nearest neighbour (KNN)

This is considered to be the best classifier in the family of nonparametric method with a good scalability. In leukemia detection kNN=1 is considered to classify between blast cells and normal lymphocytic cells [25, 27]. Bhattacharjee and Saini [25] in their research in diagnosing ALL have discovered that KNN is the best classifier that produced high specificity and also have the lowest computational complexity which has produced a specificity of 95.23%.

2.2.5. Ensemble of classifiers (EOC)

Ensemble methods are machine learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions [34]. EOC improves of the performance of individual classifiers. The ultimate goal of classification result integration algorithms is to generate more certain, precise and accurate system results. But EOC possess some limitations also such as increased storage, increase the number of computations and decreased comprehensibility. EOC is been an efficient classification model used in leukemia diagnosis so far. An ensemble classifier system for early diagnosis of ALL has been developed by Mohapatra and group in 2014 [16]. As the results they have obtained more accuracy in EOC in comparison with other classifiers employed. Scotti and Piuri [28] have used ensemble of classifiers on their research done in Morphological Classification of Blood Leucocytes by Microscope Images. The classification accuracy has been tested and a proper classifier has been chosen from a set of candidates of different classifiers.

When analysing the overall results of early researches in terms of the classification techniques used, it was understood that K-Nearest Neighbor, Support Vector Machines (SVM) and using Ensemble of classifiers have resulted better results than other identified classification metrics. The identified precipitate is summarized as in Table 2. The classification technique and the feasibility of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

diagnosing ALL has been evaluted and has been presented in Table 2 as a good or weak classifcation algorithm to employ in the developing solution.

Table 2: Summary on Classification Techniques.

Classification Technique	Good	Weak
KNN	✓	
SVM		X
ANN		X
NB		X
RFBN		X
CART		X
EOC	✓	

The accuracy of the ALL diagnosing applications incredibly depend on the segmentation and classification methods used by the researchers. As per the literature review it was revealed that segmentation using morphological operations has produced good results and in image classification, K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and using Ensemble of Classifiers (EOC) has given more accurate results. Most of the research groups have tried on detecting leukemia in isolated lymphoblasts. Recognition and splitting of overlapping red blood cells (RBC) with WBC has yet been a challenging issue in diagnosing leukemia. Furthermore, blood films lead to quality devastation when storing them for a long period. When blood films are being transported to foreign countries and long distances the quality of them degraded and important features of them get lost. Current researches have not yet been extended to detect this effect in diagnosing leukemia. Researches must be proposed in order to achieve these challenging issues which are the current live problems for the hematologists in diagnosing leukemia.

3. METHODOLOGY AND EXPERIMENTAL DESIGN

An automated diagnosing application would be a beneficial tool in diagnosing of ALL in peripheral blood samples efficiently and accurately. This section presents the methodological approach and the system design of the proposed solution developed in order to address the identified research limitations in ALL diagnosis. Ethical clearance regard to this research has been obtained prior to the initiation of study from the Ethical Review Committee (ERC) of Faculty of Medicine, Kotelawala Defence University, Sri Lanka. The basic method proposed for diagnosis proposed can be divided into steps as in Figure 2.

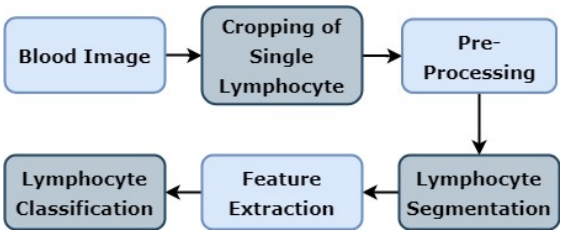


Figure 2: Basic Diagram of the System

The input for the system is the Leishman's stained blood slide image, and the lymphocytes in the image are cropped and individuated manually. Firstly, in the pre-processing module, the image acquisition noise and background non-uniformities are removed. Secondly, image segmentation is performed using proper segmentation techniques. This is done using four consecutive steps of background removal, detection of overlapped cells, separating the lymphocytic cell and separating the nucleus region which has been described in this paper. In the feature extraction module, various morphological features are being sorted differently using the segmented regions of the lymphocytic cell and the nucleus. Combining the features of both cell and nucleus, some new features are also calculated. In the lymphocyte classification module, the tested cells are labelled as blast or normal using the implemented supervised learning classification method. The block diagram of the proposed algorithm is shown in Figure 3.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

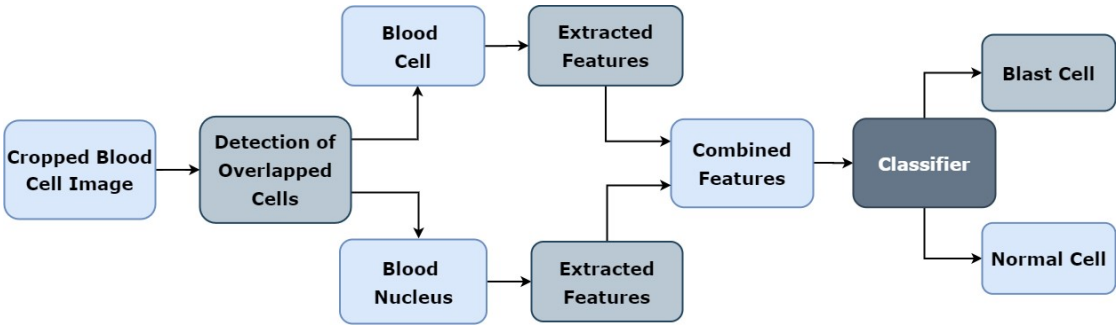


Figure 3. Block Diagram of the Proposed Algorithm

3.1. Image Acquisition

The inputs for this automation process are microscopic images obtained from peripheral blood films which stained by Leishman's that has been obtained from Department of Pathology Clinic at Faculty of Medicine, University of Colombo, Sri Lanka. All the obtained images are affected from B-ALL precursor which is a major type of ALL. The images are captured from two different camera sources as Huawei GR5 2017 smartphone camera and Canon camera in the same lightning conditions, resolution and magnification. The slides are placed under a MicroTech XSZ-N207 microscope in X100 magnification. 142 of the chosen images are taken into the experiment.

3.2. Image Pre-Processing

Pre-processing is essential as normal images consist of excessive staining and shadows. Image enhancement, which is used to bring out the image details that are obscured is the main task of this stage. Following three main tasks are performed in this stage.

3.2.1 FastNlMeansDenoisingColoured technique

This is done to remove noise and excess blurriness that is presented in coloured blood images.

3.2.2 Edge enhancement

Done in order to sharpen the image by cleaning the cell/cell segments in the boundary of the blood images.

3.2.3 RGB splitting

The RGB image is split in to three channels; green, red and blue in order to identify the red blood cells and white blood cells separately.

3.2.4 Removing the green channel

Green channels are mostly sensitive to red blood cells. Thus, it is removed from the image in this step. After removing green channel, red and blue channels are merged.

3.3. Image Segmentation

Segmentation process partitions an image to its constituent segments or objects known as pixels. This locates objects and boundaries (curves and lines etc.) of images and modifies the representation of an image into somewhat that is more meaningful and easier to analyze. This is a crucial step as the following feature extraction and classification results are much related with the result of the segmentation module. In this stage the following four steps are under gone.

3.3.1 Background removal

In this stage, canny filter is first used to reconstruct the border of the cells present in the image. Then morphological operation 'dilation' is done using a prepared structuring element. Then 'closing' is done. Combing the images obtained from dilation and closing, a new image is obtained. Next, threshold to Zero and Inverted thresholding is performed to the image obtained from pre-processing. Then the resulted image is combined with the new image obtained from morphological segmentation and the background is now removed.

3.3.2 Detection of overlapped cells

The overlapped cells of the image are detected using a combination of grab cut algorithm and circles detection algorithm.

3.3.2.1 Grab cut algorithm

Grab Cut is an image segmentation method based on graph cuts. Grab Cut Algorithm has been used in this context for separating the background in the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

overlapped cell regions. Here firstly, a user-specified bounding box around the blood plasma region to be segmented was initiated. Then, the color distribution of the target nucleus and the background plasma were estimated using a Gaussian mixture model.

3.3.2.2 Circles detection algorithm

An advancement of Circle Detection Algorithm has been developed for detecting overlapped cells. Thresholding, closing, template matching and finding the local maxima were carried out in this step.

3.3.3 Isolating the lymphocytic cell

In the resulted image, the largest contour area is considered to be the area of the cell region. The image is then subjected to a combination of binary thresholding and Otsu's thresholding and a binary image of the cell is produced. The total blood cell's binary image is now ready for feature extraction.

3.3.4 Isolating the nucleus

In this step, firstly the intensity of the original cropped blood image is increased such that only the nucleus will be visible in the image. Then thresholding is done in order to separate the nucleus. Here a combination of binary thresholding and Otsu's thresholding is done. Then the nucleus region is segmented by subjecting the image to the background removal step described earlier. Then the segmented nucleus is converted to binary and it is now ready for feature extraction. The image processing techniques listed in Image processing and segmentation stages have been used after applying and testing the image visibility and processing functionality of them with the acquisitioned blood images.

3.4. Feature Extraction

In feature extraction, the acquired data from the image is transformed and labelled to a particular set of features, which is going to be used for further classification. The binary equivalent images produced by the segmentation technique of blood cell and cell nucleus are used to extract those morphological features. Using the extracted features of blood cells and nucleus, combined features also have been acquired. The features extracted have been explored in Table 3. Feature parameters were gathered with respect to four categories; colour features, geometric features, texture features and statistical features as shown in Table 3 in detailed. As for the color features; mean color values of blood cell, cytoplasm and nucleus were extracted. Geometric features; area, perimeter, circularity, diameter, roundness, compactness, form factor, major axis length, minor axis length, convexity, solidity, ratio of area of cytoplasm to nucleus and ratio of area of nucleus to cell were extracted with respect to the cell and the nucleus. Entropy, energy, correlation, homogeneity was extracted features regard to the Texture of the cell and nucleus. Several statistical parameters; skewness, variance, mean, gradient matrix were extracted for the cell and nucleus separately. The final feature set includes 20 features including; cell area, cell perimeter, cell circularity, cell diameter, cell roundness, cell compactness, nucleus area, nucleus perimeter, nucleus circularity, nucleus diameter, nucleus roundness, nucleus compactness, nucleus form factor, nucleus to cell ratio, convexity, solidity, major axis length, minor axis length, ratio of area of cytoplasm to nucleus and ratio of area of nucleus to cell.

Table 3: Parameters Obtained in Feature Extraction.

Feature	Parameters Extracted
Colour features	Mean colour values
Geometric features	Area, Perimeter, Circularity, Diameter, Roundness, Compactness, Form Factor, Major axis length, Minor axis length, Convexity, Solidity, Ratio of area of cytoplasm to nucleus, Ratio of area of nucleus to cell.
Texture features	Entropy, Energy, Correlation, Homogeneity etc.
Statistical features	Skewness, Variance, Mean, Gradient matrix etc.

Contour detection was applied in feature extraction stage for the binary images of cell and nucleus resulted in segmentation. Generally, contours are curves that join all the continuous points along the boundary that have same color or intensity. This concept is very useful for shape analysis and object detection and recognition which was used as guideline for feature extraction step. The features of final dataset are extracted as follows. The features were gathered separately for cell and nucleus.

Area- Area is the total number of non-zero (white) pixels available within the image region. To calculate this, the contours were selected, sorted according to area, the largest contour of the segmented cell was taken and contour area was calculated using OpenCV function;

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

cv2.contourArea and determining its bounding rectangle.

Perimeter- Perimeter is the distance (major and minor both) between successive boundary pixels. It was obtained by using *cv2.arcLength* function on the first largest contour.

Circularity- This was calculated by area and perimeter.

$$Circularity = (4 * pi * Area / Perimeter^2)$$

Diameter- Was calculated using *cv2. minEnclosingCircle* function on the first largest contour.

Roundness- Is a feature representing the degree to which a shape is round.

$$Roundness= (4 * Area) / pi * Diameter*Diameter$$

Compactness- It is a numerical measure representing the degree to which a shape is compact.

$$Compactness = ((4/pi) *Area)^{1/2}/Maximum\ Diameter$$

Form Factor- It is calculated by area and circularity of both blood cell and cell nucleus.

$$Form\ Factor = ((CN\ Area / BC\ Area) * (BC\ Circularity / CN\ Circularity)$$

Where, CN = Cell Nucleus; BC = Blood Cell;

Major axis length- Length in pixels of the major axis of the ellipse containing the nucleus.

Minor axis length- Computed as the length in pixels of the minor axis of the ellipse containing the nucleus.

Convexity- When calculating the convexity, firstly a convex hull was obtained using *cv2.convexHull* (firstlargestcontour_nuclues) function. Then *cv2.contourArea* function was used to obtain the convexity.

Solidity- *Area / Convex Area*

Ratio of area of cytoplasm to nucleus.

Ratio of area of nucleus to cell.

The selected feature set was visualized using Seaborn and matplotlib plot and statistically analyzed, which has been explained under evaluation section.

3.5. Image Classification

The features extracted in feature extraction stage were directed to the image classification stage which employs a supervised learning classification. For classification, K-Nearest Neighbour (k-NN) algorithm is used. k-NN algorithm is a non-parametric method used in pattern recognition for classification and regression. This classifier has been used in classification of the cells as ALL affected or healthy. In k-NN classification, the output is generally a class membership. In this context, the data points were classified by a majority vote of its neighbors, with the points being assigned to the class most common among its k nearest neighbors. The label 1 has been assigned for detected acute leukemic cells and label 0 has been assigned for healthy cells respectively. Statistical data visualization on classified results was done using Seaborn and matplotlib plot analysis.

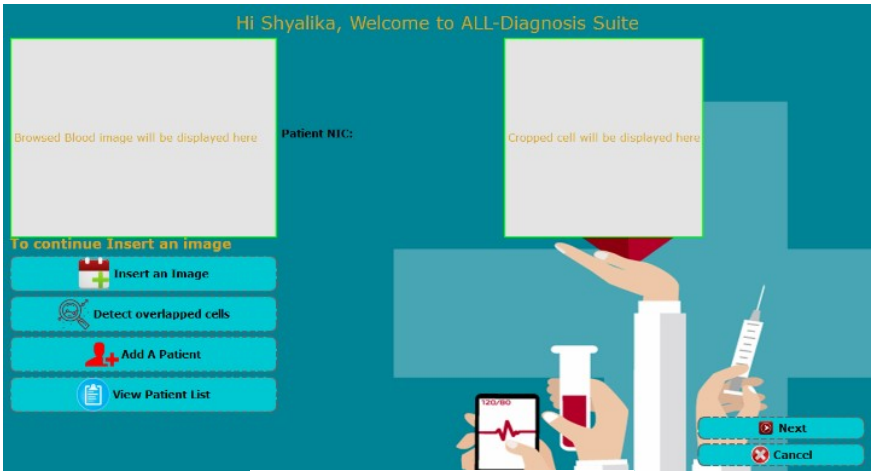


Figure 4: ALL Diagnosing Window

Journal of Theoretical and Applied Information Technology
30th September 2020. Vol.98. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EVALUATION AND RESULTS

There are two main objectives that this research is focused on. One is to develop an algorithm to compare the image qualities of purely isolated cells and overlapping cells. Second one is to detect the quality devastation in blood films in terms of storing them for prolonged period. The proposed algorithm was implemented using python programming language using the OpenCV package for python. Figure 4 depicts the design of a user interface in the GUI suite done using Qt. Figure 8 and Figure 9 are the results that were obtained for one instance in the feature extraction respectively for the cell and the nucleus. Figure 5 presents an image of the microscopic image used for testing. Figure 6 elaborates a results of an isolated lymphocyte image used for segmentation stage of the cell and nucleus. In the visualization, cell segmentation steps and nucleus segmentation steps are shown separately.

In cell segmentation steps, the pre-processed image shows the result of the steps FastNlMeansDenoisingColoured technique, Edge enhancement, RGB Splitting and removing the green channel respectively. Next canny, dilation, closing and combination of dilated and closed images, thresholding for pre-processed image and binary image of the cell has been shown. In nucleus segmentation step firstly, original image is shown. Then the thresholding mask image shows the result of the combination of the binary and Otsu's segmentation performed. Next canny, dilation, closing and combination of dilated and closed images, thresholding for pre-processed image and binary image of the cell has been shown.

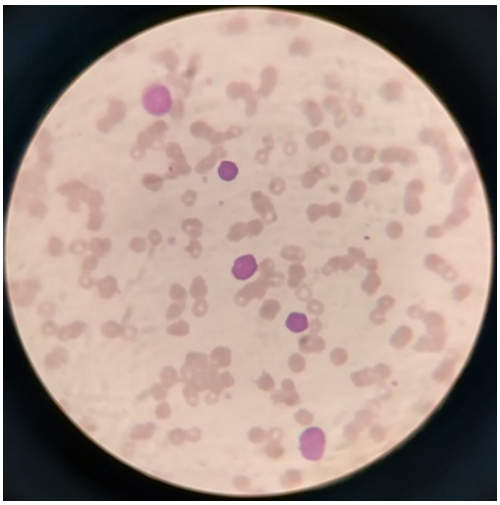


Figure 5: Sample Microscopic Image used for Testing.

In the feature extraction module, chosen geometric features that are appropriate for the data set with regard to the cell and nucleus had been extracted separately. The feature set includes 20 features including; cell area, cell perimeter, cell circularity, cell diameter, cell roundness, cell compactness, nucleus area, nucleus perimeter, nucleus circularity, nucleus diameter, nucleus roundness, nucleus compactness, nucleus form factor, nucleus to cell ratio, convexity, solidity, major axis length, minor axis length, ratio of area of cytoplasm to nucleus and ratio of area of nucleus to cell. Exploratory data analysis was performed to analyze relation between each and every feature variable. Seaborn pairs plotting was used to discover patterns in the feature dataset by considering distribution of single variable (univariate analysis) and relationships between two variables (bivariate analysis). The plotted results on a selected sample for a selected feature set is shown in Figure 7.

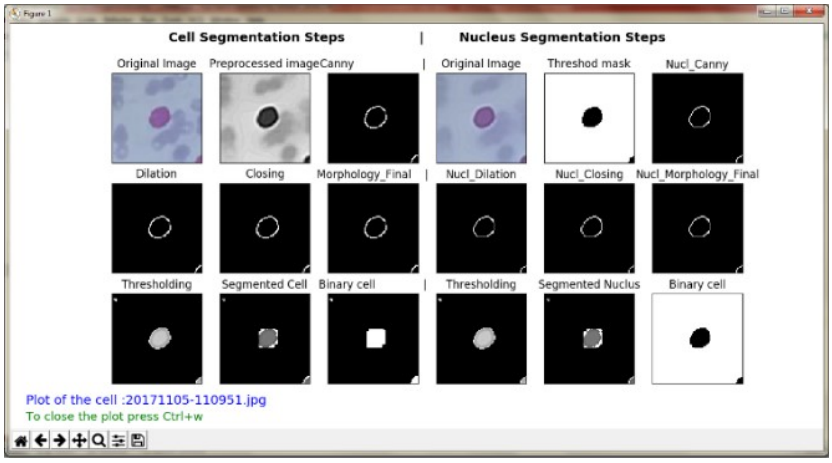


Figure 6: Segmentation Results of the Cell and Nucleus

Journal of Theoretical and Applied Information Technology
30th September 2020. Vol.98. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

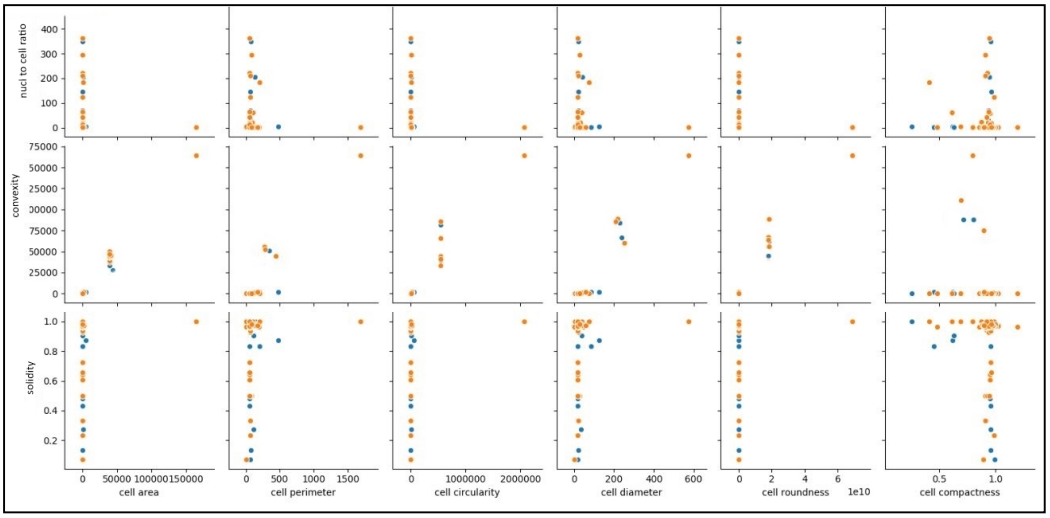


Figure 7: Exploratory Analysis on Selected Sample

The features extracted with respect to a one lymphocytic cell and nucleus are depicted in Figure 8 and Figure 9. The features have been saved in a csv file and continued for use in classification stage.

The proposed algorithm has been tested and developed to detect features of overlapping cells and to detect the quality devastation that occurs in old blood films when they are kept for 3-6 months. In detecting overlapping cells, firstly the overlapping property was identified from the acquisitioned images. Then overlapped cells were isolated using the implemented algorithm, then they were treated as separate individual cells and continued for ALL diagnosing algorithm proposed. For convenience, overlapping cells that have two overlapped cells were taken into consideration in the first phase. These two blood cells were observed in two layers in the blood image when the 3D bilateral layers were onserved. The summary of overlapped cells and isolated cells contributed to the evaluation is added to Table 4. Figure 10 depicts how the overlapping cells were detected and identified in the developed algorithm. In order to provide a solution for the quality devastation of blood films, images of blood films were taken in 2 attempts keeping a time gap of 6 months in between. The algorithm was modified taking normalized values that match with both the fresh and old blood films.

The details of the blood images used are as follows in Table 4. The microscopic images were obtained from two ALL affected patients and one healthy person respectively in consecutive two rounds as shown in Table 4.

{'m00': 495.0, 'm01': 19368.333333333332, 'nu03': 0.000574044278723185, 'm03': 31952400.1, 'mu03': 3129.3806308470666, 'm30': 99167335.15, 'mu12': 2193.0417784254532, 'mu02': 19563.396745230188, 'nu30': -0.00031116751484171505, 'nu02': 0.07984245177116697, 'm02': 777406.5, 'm10': 28630.166666666664, 'm20': 1675481.0, 'mu30': -1696.317914813757, 'mu11': -120.53563411883079, 'mu20': 19548.791189674754, 'nu20': 0.07978284334118867, 'm12': 44956957.71666667, 'm11': 1120119.0833333333, 'nu11': -0.0004919319829357444, 'nu12': 0.0004022850635351724, 'nu21': -0.0005909917907598672, 'm21': 65540965.28333333, 'mu21':-3221.7693504535127}

Cell features
Area: 495.0
Perimeter: 83.59797883033752
Circularity: 6222.857142857143
Diameter: 26
Roundness: 425880.0
Compactness: 0.965376953693164

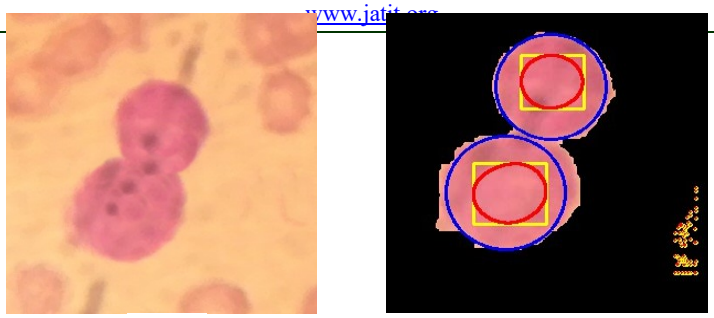
Figure 8: Feature Extraction Results of Cell

{'m00': 449.0, 'm01': 17522.0, 'nu03': 0.0006957435707228384, 'm03': 28553333.0, 'mu03': 2972.111199848354, 'm30': 90211112.25, 'mu12': -435.8012352756923, 'mu02': 15937.897178916144, 'nu30': 0.0002335588959071145, 'nu02': 0.07905663751130274, 'm02': 699725.1666666666, 'm10': 26020.5, 'm20': 1524172.6666666665, 'mu30': 997.7282429337502, 'mu11': -223.05660727538634, 'mu20': 16229.63715664437, 'nu20': 0.0805037532385472 8, 'm12': 40532708.75, 'm11': 1015213.9166666666, 'nu11': -0.0011064260954825934, 'nu12': -0.0001020170132165998 6, 'nu21': -0.0007407147693901 7, 'm21': 59451055.583333336, 'mu21':-3165.7450277126627}

Nucles features
Area: 449.0
Perimeter: 79.59797883033752
Circularity: 5644.571428571428
Diameter: 24
Roundness: 329157.81818181823
Compactness: 0.9960464777602428
Form Factor: 1.0000000000000002
Nuclues to Cell ratio: 1.1024498886414253
Cytoplasm area: 46.0
Cytoplasm to nuclues ratio: 9.76086956521739
Convexity: 458.0
Solidity: 0.980349344978166

Figure 9: Feature Extraction Results of Nucleus

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) (b)

Figure 10: (a) Overlapped cells (b) Isolation of overlapped cells

Table 4: Summary of Number of Images Obtained.

Number of microscopic images obtained	Patient 1 (ALL affected)		Patient 2 (ALL affected)		Patient 3 (Healthy)	
	Isolated cells	Overlap cells	Isolated cells	Overlap cells	Isolated cells	Overlap cells
First Round	350	100	300	180	420	220
Second Round	300	150	240	200	410	250
Total Images	650	250	540	380	830	470
ALL Affected vs Healthy	1820				1300	

The number of overlapped cell instances and isolated cell instances obtained from the same are included in Table 4. Total number of 2184 images (ALL-1274, Healthy-910) were used as training dataset and 936 (ALL-546, Healthy-390) images were reserved for testing purpose as represented detailed in Table 5. Evaluation matrix / confusion matrix was built to evaluate the performance of the classification model. The evaluation matrix resulted from the evaluation is presented in Table 6 where the actual class and predicted class is weighed accordingly. The results were categorized into four partitions; correct diagnosis of healthy, error diagnosis of healthy, error diagnosis of ALL and correct diagnosis of ALL. Table 7 includes the legend of the evaluation matrix.

Table 5: Number of Images used in Training and Testing.

Training & Testing Datasets	Total Number of images used	ALL affected	Healthy
Training Data set	2184	1274	910
Testing Data set	936	546	390

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6: Evaluation Matrix with the Total Results Obtained.

Evaluation Matrix	Predicted Class		
Actual Class		Class=ALL	Class=Healthy
	Class=ALL	702 (TP)	52 (FN)
	Class=Healthy	52 (FP)	130 (TN)

Table 7: Legend of Evaluation matrix.

Name	Description
TN	True Negative-correct diagnosis of healthy
FP	False Positive- error diagnosis of healthy
FN	False Negative- error diagnosis of ALL
TP	True Positive- correct diagnosis of ALL

As per the evaluation results, it was able to achieve an average accuracy of 88.8% in diagnosing ALL. The accuracy was calculated as ratio between the total number of correctly diagnosed ALL or healthy results and the total samples tested.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \times 100$$

$$= 88.8\%$$

The error rate/misclassification rate was calculated as the ratio between the total incorrectly diagnosed samples to the total tested samples as follows.

$$\text{Misclassification Rate} = \frac{FP + FN}{TP+TN+FP+FN} \times 100$$

$$= 11.11\%$$

Overall, the algorithm holds 11.11% of misclassification rate. Accordingly, there seem to have some place for improve, when considering the quality of blood images obtained, environment conditions, normalizing indicators etc.

To get the value of precision, the total number of correctly diagnosed ALL effected positive samples was divided by the total number of predicted positive examples. High Precision resulted with a less number of error diagnosis of healthy, indicates the results labelled as positive ALL diagnosed is indeed positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$= 0.93$$

The Recall was defined as the ratio between the total number of correctly diagnosed ALL effected positive examples to the total number of positive samples obtained by the addition of correctly diagnosed ALL affected and error diagnosis of ALL. This resulted in considerable high recall (0.93%) which results in the success of the classifier model.

$$\text{Sensitivity / Recall} = \frac{TP}{FN+TP}$$

$$= 0.93$$

The end product of the research comprises dual components interconnected; a Python based Qt GUI enabled development suite and a PHP based web application. Python suite performs main targeted backend functionalities and web application that serves hematologists, doctors and patients to perform utility functions. The end system includes modules for login and authentication, user registration, main window, ALL diagnosing, patient information recording, records viewing, diagnosed results plotting, database functionalities, report generation and diagnosed reports sending for patients via email.

5. CONCLUSION

The developed system has been extended to have good results in automatic diagnosis of the disease in the acquired human blood samples. It has obtained an average accuracy of 88.8% in diagnosing ALL in human blood samples. The proposed algorithm can also be further developed to detect the granules and intra cellular components inside the cell. Using more than one classifier in the aim of increasing the accuracy of the proposed algorithm has been identified as a further work in this research. As per the statistical data published in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

future spreading of a cancer like leukemia in the world, automat ion procedures to detect leukemia has become an urgent need. Governments, especially in developing countries like South Asia would find these automatic leukemia diagnosing systems as cost effective solutions to implement in hospitals.

ACKNOWLEDGEMENTS

The research is financially supported by Mobitel (Pvt) Ltd, Sri Lanka. Authors would greatly appreciate and acknowledge their contribution, whose dedicated assistance for the publication that should be highly recognized.

The corresponding author would acknowledge the supervisors Mr. PPNV Kumara and Dr. Darshana Kottahachchi whose insight and expertise that greatly assisted the research. Special thanks goes to all the lecturers and staff of Faculty of Computing and Department of Medical Laboratory Sciences of Faculty of Allied Health Sciences at General Sir John Kotelawala Defence University, Sri Lanka whose loyal support for the research project was highlighted in all stages of the research. Authors are also grateful to the staff of Department of Pathology Clinic at Faculty of Medicine, University of Colombo, Sri Lanka who supported by providing research materials and Faculty of Medicine at General Sir John Kotelawala Defence University, Sri Lanka for granting ethical approval for the research.

AVAILABILITY OF DATA AND MATERIAL

The dataset used to support the findings, source code and the executable file developed in this research project are available from the corresponding author upon request.

REFRENCES

[1] Catovsky D, Hoffbrand AV. Acute leukemia. In: Hoffbrand AV, Lewis SM, Tuddenham EDG (eds), “Postgraduate Haematology”, 4th edition, Arnold: London, 2001, pp 373–404.

[2] Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnic HR, Sultan C, “Proposals for the Classification of Acute Leukemia”, Br J Haematology, 33: 451-8,1976.

[3] Hughes-Jones NC, Wickramasinghe SN, Hatton CSR, “Lecture notes on Hematology”, Seventh Edition, Wiley Blackwell Publishing 2004

[4] Purohit SS, “Biotechnology Fundamentals and Applications”, Third Edition,2000.

[5] Cui J, Wang J, He K, Jin B, Wang H, Li W, Kang L, Hu Li MH, Yu M, Shen B, Wang G and Zhang X, “Proteomic analysis of Human Acute Leukemia Cells: Insight into Their Classification, Clinical Cancer Research” Volume. 10, October 15,2004, pp. 6887 – 6896.

[6] Scholl I, Aach T, Deserno TM, Kuhlen T, “Challenges of Medical Image Processing.” Computer Science - Research and Development, Volume 26, Issue 1-2, 2011, pp. 5–13, doi:10.1007/s00450-010-0146-9.

[7] Hegadi RS, “Image Processing: Research Opportunities and Challenges.”, National Seminar on Research in Computers (NSRC), 13-12-2010, Bharathiar University, Coimbatore, India December, 2010.

[8] Dodandeniya JMDGCM, Kumara PPNV, and Kulasekara DMR, “Automated Blood Counter (ABC): An Image Processing Solution”, Proceeding of International Symposium of General Sir John Kotelawala Defence University, September, 2016

[9] Vaghela HP, Modi H, Pandya M, Pandya M, “Leukemia Detection using Digital Image Processing Techniques”, International Journal of Applied Information Systems (IJAIS)–ISSN 2249-0868, Volume 10, Issue 01, November, 2015.

[10] Patil BG, Jain SN, “Cancer Cells Detection Using Digital Image Processing Methods.” International Journal of Latest Trends in Engineering and Technology (IJLTET), Volume 3, Issue 4, March, 2014.

[11] Mahaja S, Golait SS, Meshram A, Jichlkan N, “Review: Detection of Types of Acute Leukemia”, International Journal of Computer Science and Mobile Computing, Volume 3 Issue 3, March- 2014, pp. 104-111.

[12] Viswanathan P, “Fuzzy C Means Detection of Leukemia Based on Morphological Contour Segmentation”, Procedia Computer Science 58 (2015) 84 – 90, Second International Symposium on Computer Vision and the Internet (VisionNet’15), Volume 58, pp. 84–90. doi: 10.1016/j.procs.2015.08.017.

[13] Mohapatra S, Samanta SS, Patra D, Satpathi S, “Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection”, IEEE

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

International Conference on Devices and Communications (ICDeCom), 2011, pp. 1–5.

[14] Scotti F, "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images", IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Italy, July, 2005, pg. 96–101.

[15] Mitra S, Pedrycz W, Barman B, 2010. "Shadowed c-means: Integrating fuzzy and rough clustering". Pattern Recognition, Volume 43, Issue 4, pg. 1282–1291, doi: 10.1016/j.patcog.2009.09.029

[16] Mohapatra S, Patra D, Satpathy S, An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images", Neural Computing and Applications, Volume 24, Issue 7-8, 2014, pg.1887–1904. doi:10.1007/s00521-013-1438-3

[17] Olivera JV, Pedrycz W (Editors), Winkler R and Klawonn F and Kruse R (Writers), "Advances in fuzzy clustering and its applications", Wiley, Chichester, 2007.

[18] Cheng HD, Jiang XH, Sun Y and Wang JL, "Color Image Segmentation: Advances & Prospects", Department of Computer Science, Utah State University, Logan, 2006.

[19] Nor Hazlyna H, Mashor MY, Mokhtar NR, Aimi Salihah AN, Hassan R, Raof RAA, Osman MK, "Comparison of Acute Leukemia Image segmentation using HSI and RGB color space". 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010) (IEEE),2010, pp. 749–752, doi:10.1109/ISSPA.2010.5605410

[20] Singhal V, Singh P, "Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia", 2014 Twentieth National Conference on Communications (NCC), IEEE, 2014, pp. 1–5.

[21] Halim NHAbd, Mashor MY, Nasir ASA, Mokhtar NR, Rosline H, "Nucleus segmentation technique for acute leukemia", 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications (CSPA), IEEE, 2011, pp. 192–197.

[22] Fatma M, Sharma J, "Identification and Classification of Acute Leukemia using Neural Network", 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), IEEE, 2014, pp.142–145.

[23] Rejintal A, Aswini N, "Leukemia Cancer Cell Detection using Image Processing", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 4th National Conference on Design Innovations for 3Cs "Compute-Communicate-Control", Volume 5, Special Issue 6, Dept. of ECE, MVJ College of Engineering, Bangalore-560067, India, July 2016.

[24] Agaian S, Madhukar M, Chronopoulos AT, "Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images", IEEE Systems Journal, Volume 8, Issue 3, 2014, pp. 995–1004, doi: 10.1109/JSYST.2014.2308452

[25 Bhattacharjee R, Saini LM, "Robust Technique for the Detection of Acute Lymphoblastic Leukemia", 2015 IEEE Power, Communication and Information Technology Conference (PCITC), IEEE, Siksha 'O' Anusandhan University, Bhubaneswar, India, 2015, pp. 657–662.

[26] Raje C, Rangole J, "Detection of Leukemia in Microscopic Images using Image Processing", 2014 International Conference on Communications and Signal Processing (ICCSP), IEEE, India, April 3-5, 2014, pp. 255–259.

[27] Shyalika JKC, Kumara PPNV and, Kottahachchi DU, "An Image Processing Application for Diagnosing Acute Lymphoblastic Leukemia (ALL)", 10th International Research Conference, General Sir John Kotelawala Defence University, Sri Lanka. General Sir John Kotelawala Defence University, Sri Lanka, August, 2017, p.7. Available at: http://ir.kdu.ac.lk/handle/345/1695

[28] Piuri V, Scotti F, "Morphological Classification of Blood Leucocytes by Microscope Images", CIMSA 2004 - IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, IEEE, Boston, MA. USA, 14-16 July 2004, pp. 103–108.

[29] Rege MV, Dr. Gawli BW, "Detection of Leukemia in Human Blood Sample through Image Processing: A Review", International Journal of Modern Trends in Engineering and Research, Volume 02, Issue 10, October 2015, pp.363-369.

[30] Patel N, Mishra A "Automated Leukaemia Detection Using Microscopic Images", Procedia Computer Science, Volume 58, Second

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

International Symposium on Computer Vision and the Internet (VisionNet'15), pp. 635–642, doi: 10.1016/j.procs.2015.08.082.

[31] Faivdullah L, Azahar F, Htike ZZ, Naing WYN, "Leukemia Detection from Blood Smears". Journal of Medical and Bioengineering, Engineering and Technology Publishing, Volume 4, No. 6, December 2015, pp. 488–491, doi:10.12720/jomb.4.6.488-491.

[32] Mohapatra S, Patra D, Kumar S, Satpathy S, "Lymphocyte Image Segmentation using Functional Link Neural Architecture for Acute Leukemia Detection". Biomedical Engineering Letters, Volume 2, 2012, pp. 100–110, doi:10.1007/s13534-012-0056-9.

[33] Serbouti S, Duhamel A, Harms H, Gunzer U, "Image Segmentation and Classification Methods to Detect Leukemias", Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Volume 13, Issue 01, France, 1991, pp. 260–261.

[34] Dietterich TG, "Ensemble Methods in Machine Learning", Oregon State University, Corvallis, Oregon, USA, 2011.