2019 6th International Conference on Control, Decision and
Information Technologies (CoDIT'19) | Paris, France / April 23-26, 2019

# Automatic Acute Lymphoblastic Leukemia Detection and Comparative Analysis from Images

Md. Nuruddin Qaisar Bhuiyan[1], Shantanu Kumar Rahut[2], Razwan Ahmed Tanvir[3] and Shamim Ripon[4]

***Abstract*— In this era, surrounded by numerous technologies, medical sector has seen a lot of advancement through implementing various autonomous systems to identify different types of diseases. In this paper, a framework for identification of Acute Lymphoblastic Leukemia from the microscopic image of white blood cell is proposed. Microscopic images are at first carefully preprocessed to prepare them for classification. In addition, four different machine learning algorithms, namely, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) are applied and respective results are analyzed to provide a comparison between these algorithms in terms of different performance metrics. After a thorough comparison, it is observed that the SVM works well to classify and identify the Acute Lymphoblastic cell which is responsible for Leukemia cancer.**

## I. INTRODUCTION

Leukemia is a type of cancer which begins in a cell in the bone marrow. The bone marrow cells then gradually change and take the form of leukemia cell. Leukemia cells gradually increase and suppress the development of normal blood cell. There are three components of blood (i) Red Blood Cell or Erythrocytes (RBC) (ii) White Blood Cell or Leucocytes (WBC) and (iii) Platelets. Leukemia can be detected only by analyzing the WBCs. So, focus of this study is only on the white blood cells. Leukemia can be classified into two separate classes,

1) *Acute Leukemia*: The affected WBCs do not act like normal blood cells and gradually increase in number.
2) *Chronic Leukemia*: The affected WBCs act like normal blood cells, also gradually increase in number.

There are subdivisions of leukemia based on the kind of blood cell affected. These are,

1) *Lymphoblastic* or Lymphocytic Leukemia
2) *Myeloid* or Myelogenous Leukemia

This paper focuses only on the Acute Lymphoblastic Leukemia (ALL) which is the most common kind of leukemia. In ALL, bone marrow makes too many immature lymphocytes. Symptoms of ALL include tiredness, fever, enlarged lymph nodes, bone pain etc. Like any other acute leukemia, ALL can become fatal in several weeks or months, if there is no proper treatment. Acute lymphoblastic leukemia is acute leukemia of lymphoblastic type, one of the two major categories of acute leukemia, primarily affects young children.

Stem cells can evolve into myeloid stem cells or lymphoid stem cells. Myeloid stem cells evolve into myeloid blasts and lymphoid stem cells evolve into lymphoid blast [17]. Blast cells are premature circulating blood cells such as neutrophils, monocytes, lymphocytes and erythrocytes. Blasts are usually found in low numbers in the bone marrow. They are not usually found in significant numbers in the blood. Blast cells are premature form of white blood cells. Acute Lymphoblastic leukemia can be detected from infected white blood cell. All other cells other than the blast cells are known as non-blast cells.

The diagnostic process of detecting ALL can be done through the inspection of microscopic image of white blood cell. This inspection is performed by the experts to identify ALL. But the expansion in image processing and other technologies, a new door for automatic ALL detection from microscopic image has been opened. This process is inexpensive and very fast. Besides, standard accuracy can also be maintained.

Most patients with ALL have too many immature white cells called lymphoblasts (or just blasts) in their blood, and not enough red blood cells or platelets. Lymphoblasts are not normally found in the blood, and they do not function like normal, mature white blood cells. These blast cells are premature circulating blood cells such as neutrophils, monocytes, lymphocytes and erythrocytes and they are usually found in low numbers in the bone marrow. They are also not usually found in significant numbers in the blood. ALL can be detected from infected white blood cell.

Several studies on automatic leukemia detection have been performed. A similar work as ours has been done recently on the ALL-IDB1 database in [22]. The work proposes an approach for detecting acute lymphoblastic leukemia. The approach includes removal of background using Zack algorithm, feature extraction using Hough transform, separating grouped cells using distance transformation of watershed segmentation. The authors claimed that the algorithms they used, failed to produce accurate results in some scenarios. In comparison, we have taken a different mechanism for image processing and classification. Most importantly our proposed framework is able to produced better result than that work where the achieved highest accuracy is 99.05% by using support vector machine classifier.

The authors in [2], [6], [8]–[10], [21] proposed automatic leukemia detection framework based on image processing

[1]Md. Nuruddin Qaisar Bhuiyan, [2]Shantanu Kumar Rahut and [3]Razwan Ahmed Tanvir graduated from the Department of Computer Science and Engineering, East West University, Bangladesh.
[1]qaisar.ewu.cse14@gmail.com
[2]shantanurahut@gmail.com
[3]razwantanvir8050@gmail.com
[4]Shamim Ripon is with the Department Computer Science and Engineering, East West University, Bangladesh dshr@ewubd.edu

978-1-7281-0521-5/19/$31.00 ©2019 IEEE

whereas WBC image segmentation based on fuzzy morphology is proposed in [5] and images of RBC is analyzed in [2], [12]. On the other hand, authors in [15] proposes a functional Link Neural Architecture for ALL detection. Statistical Texture Analysis was done to detect ALL in [16]. In [19], a high throughput screening algorithm for leukemia cells is proposed. Many studies are based on the performance evaluation of leukemia cell detection [4], [20]. Classification system is developed based on morphological features in [18], using k-means in [14] and using SVM in [23]. Along with these, Raman Spectroscopy is applied in [13] and Dynamic Short Distance Pattern Matching Algorithm has been used in [1].

The work presented in this paper aims at comparing the accuracies of different machine learning algorithms and providing a comparative analysis of these algorithms to automatically detect Acute Lymphoblastic Leukemia. Four different algorithms, namely, Random Forest, Support Vector Machine, Logistic Regression and Decision Tree are applied in this paper. A comparative analysis of the performance of the algorithms has been done by using various performance metrics. The rest of the paper is organized as follows. Section II describes the proposed framework of the presented work. The following section demonstrates the implementation details of the work. Analysis of result is presented in Section IV. Finally, Section V concludes the work by summarizing the contribution of the work and outlining our future plan.

## II. PROPOSED FRAMEWORK

The proposed framework follows few steps from inputting image to obtaining result. At first, the input images are collected. Usually, these images are of various sizes along with various resolutions. Having uniform sized images make them suitable for any kind of experiment and analysis. Hence these images are then resized and color filtered for separating region of interest. Detecting features that are suitable for the target mining algorithm is a challenging task. After careful analysis of the images and considering our objective, feature detection and description of the images are performend by using Canny edge detector and HOG feature descriptor. Feature dimension is reduced by using Principle Component Analysis. Then for classification, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression are used. The proposed framework is illustrated in Fig. 1.

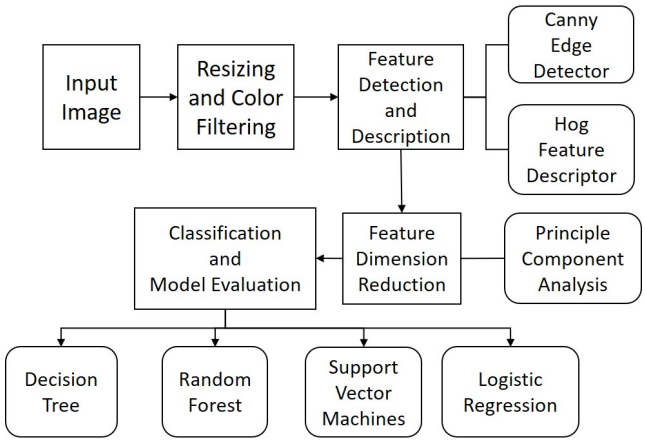


Fig. 1. Proposed Framework

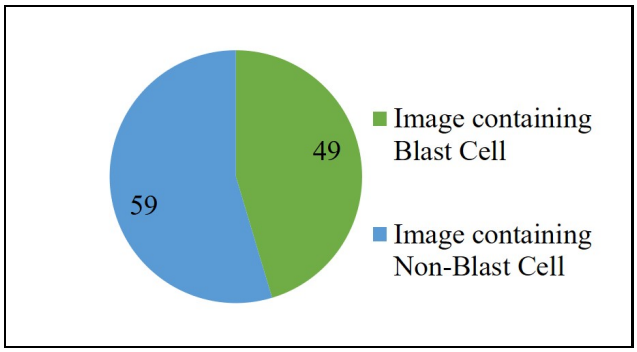


Fig. 2. ALL-IDB1 Dataset

## III. IMPLEMENTATION

### *Dataset Description*

For implementing the proposed methodology, microscopic images of white blood cell are collected from ALL-IDB1 Database [11]. This dataset is composed of 108 images. Among these image, 49 images contain blast cell and 59 images contain non-blast cell (Fig. 2). It contains about 39000 blood elements. In this dataset, lymphocytes have been labeled by expert oncologists. The number of lymphoblast among the elements is 510 [11].

Example images from the dataset have been shown in Fig. 3 to give an insight on the tissue images.

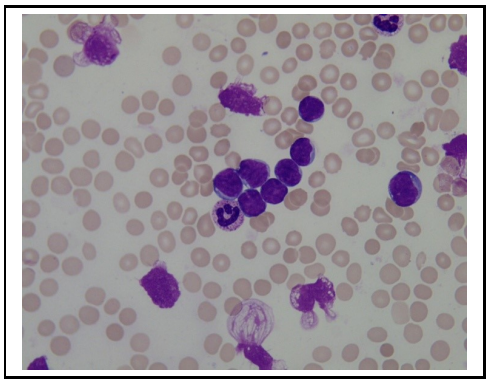

Fig. 3. Tissue image from dataset containing blast cells

Fig. 3 shows a tissue image from the dataset that contains blast cells. A non-blast cell can also be detected from the tissue image.

Fig. 4 illustrates a tissue image from the dataset that contains non-blast cells.

### *Preprocessing*

Preprocessing is a step of preparing the raw data suitable for applying various algorithms. The sample images that are collected for the experiments contain both blast and non-blast

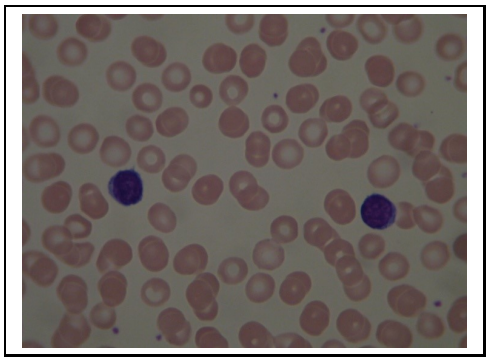

Fig. 4. Tissue image from dataset containing non-blast cells

types of cells. First of all, the images that have been collected are resized to $1000 \times 1000$ pixels. The images shown in Fig. 5 and the other images in this section are cropped from the tissue for better visualization.

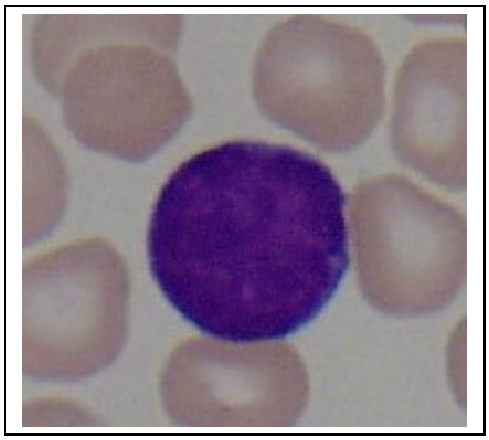

Fig. 5. Cropped microscopic image of white blood cell

*Extraction of blast and non-blast cells*

The resized ($1000 \times 1000$ pixels) images contain both blast cells and non-blast cells. For separating the region of interest from the whole image, color filtering technique has been applied. First, images of RGB formats were converted to HSV format. After that, everything outside the target color range (purple) were filtered out. Finally, the region of interest containing only the cell was found as shown in Fig. 6.

To check if the color filtering yielded the right result, bitwise logical AND of original image with the image found after color filtering was done. And the result of bitwise AND proved that, the color filtering process is a success which is shown in Fig. 7.

*Feature Extraction*

In the extracted images, the cells are of round shape. For convenience, while doing feature detection, canny edge detection algorithm is used for every image. First, Gaussian filter is used to smooth the image for removing noise. Then intensity gradients of the image are calculated. For getting rid of spurious response to edge detection non-maximus

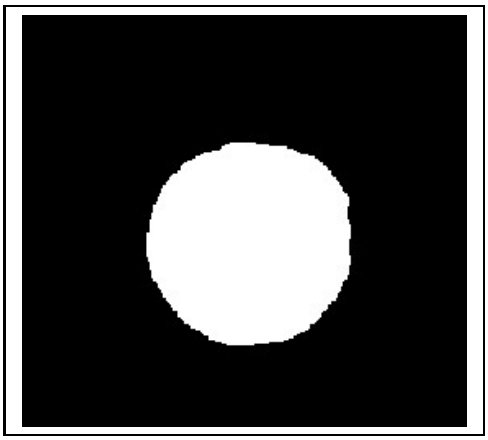

Fig. 6. Image after color filtering process

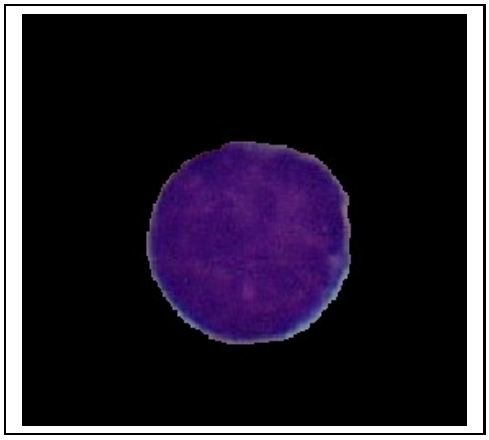

Fig. 7. Image after bitwise AND operation was done on original image and the result of color filtering

suppression is applied. Then to determine potential edges of the image, double threshold is applied. After that edges of the image are tracked through hysteresis. And for finalizing the edge detection process, weak and not connected to strong edges are suppressed [3]. The resultant image is shown in Fig. 8. In this experiment, the default parameters of sklearn of Phython have been used.

*Feature Description*

For feature description process, HOG (Histogram of Oriented Gradients) feature descriptor is used. For implementing the HOG feature descriptor on an image, at first gradient is calculated for that image. Then through Gaussian weighting and Bin assignment method a histogram is generated. After that block normalization technique is used. Finally, the descriptor of that particular image is calculated [7]. The resultant image is presented in Fig. 9. The parameter values considered for the experiments are as follows, Orientations: 9, Pixel/cell: $(10, 10)$, Cells/block: $(2, 2)$ and Method: L2-Hys block normalization.

*Dimensionality Reduction*

After applying HOG feature descriptor, each image is converted into a 1-D array called feature vector. This fea-

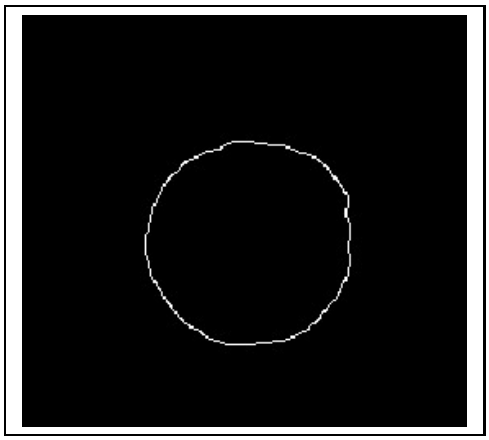

Fig. 8. Image after feature detection with canny edge detection

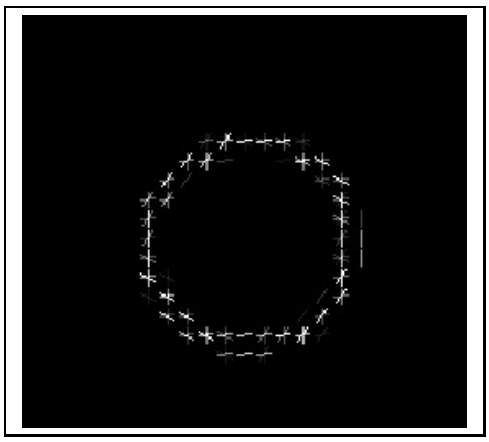

Fig. 9. Image after feature detection with canny edge detection

ture vector has $3,52,836$ dimensions which are very large; working with these huge number of dimensions may reduce the performance. So, PCA (Principal Component Analysis) dimension reduction technique is used to reduce the number of dimensions or columns of the resultant array. PCA has been used so that, the dimensions of the resultant feature vector would be reduced and consequently the performance may be improved. The classification process begins after getting the reduced feature dimension. For experimental purpose, in this work, only top ten principle components are used for classification, which means the dimension of the feature vector is 10.

*Classification and Evaluation*

For classification, Random Forest, Decision Tree, Logistic Regression and Support Vector Machine algorithms are used. Since the dataset does not contain any test set, we split the dataset into 5-folds. Among them, 4 folds are used to train and the remaining part is used to test, like the cross validation. Thus, each fold is tested and corresponding accuracy, precision, recall and f1-score are generated for each test. Then average of the results are calculated for evaluation of the model. We believe that such splitting and then evaluating each fold separately improve the accuracy of the experiments. In our experiments, information gain and entropy are used for decision tree. For random forest, $5-30$ estimators are chosen for cross-validation. In SVM, the following parameter values are used, C: $[0.0001, 0.001, 0.01, 0.1, 1, 10]$, Gamma: $[0.001, 0.01, 0.1, 1]$ and Kernel: RBF. Finally, in logistic regression the values $[0.0001, 0.001, 0.01, 0.1, 1, 10]$ are chosen by cross-validation for C_options.

## IV. RESULT ANALYSIS

In the field of disease prediction, only high accuracy percentage is not enough to choose the best model. Since the main focus of this type of classification problem i.e. disease prediction, is to correctly classify the positive classes which in this case is blast cell, having a high recall percentage can contribute to choosing the best model. As precision value and recall value has a tradeoff between them, we also calculated f1-score for each of the model, which is the harmonic average of precision and recall. Table I illustrates the result obtained from the proposed model.

TABLE I
COMPARATIVE ANALYSIS OF VARIOUS MINING ALGORITHMS

| Classifier | Precision (%) | Recall (%) | F-score (%) | Accuracy (%) |
|---|---|---|---|---|
| Random Forest (RF) | 98.4 | 98.00 | 98.00 | 98.00 |
| Support Vector Machine (SVM) | 99.20 | 99.00 | 99.00 | 99.05 |
| Logistic Regression (LR) | 97.60 | 97.00 | 97.00 | 97.18 |
| Decision Tree (DT) | 98.40 | 98.00 | 98.00 | 98.18 |

From the Table I, we could clearly define the lower limit of accuracy on ALL-IDB1 dataset. After using four different classification algorithms, it can be seen that accuracy of Random Forest classifier is 98.00%, accuracy of Support Vector Machine classifier is 99.05%, accuracy of Logistic Regression classifier is 97.18% and accuracy of Decision Tree classifier is 98.18%. The lowest accuracy that was achieved is 97.18%, by using Logistic Regression. So, the lower limit of classification accuracy on the ALL-DB1 dataset can be defined as 97.18%.

From the Table I, we could clearly define the lower limit of accuracy on ALL-IDB1 dataset. After using four different classification algorithms, it can be seen that accuracy of Random Forest classifier is 98.00%, accuracy of Support Vector Machine classifier is 99.05%, accuracy of Logistic Regression classifier is 97.18% and accuracy of Decision Tree classifier is 98.18%. The lowest accuracy that was achieved is 97.18%, by using Logistic Regression. So, the lower limit of classification accuracy on the ALL-DB1 dataset can be defined as 97.18%

From the graph in Fig. 10, it is clear that Support Vector Machine classifier gives the best performance with the accu-

racy, precision, recall and f1-score and the Logistic Regression classifier gives the worst among all used classifiers. On the other hand, though the Decision Tree has a high accuracy than Random Forest but they have similar precision, recall and f1-score.

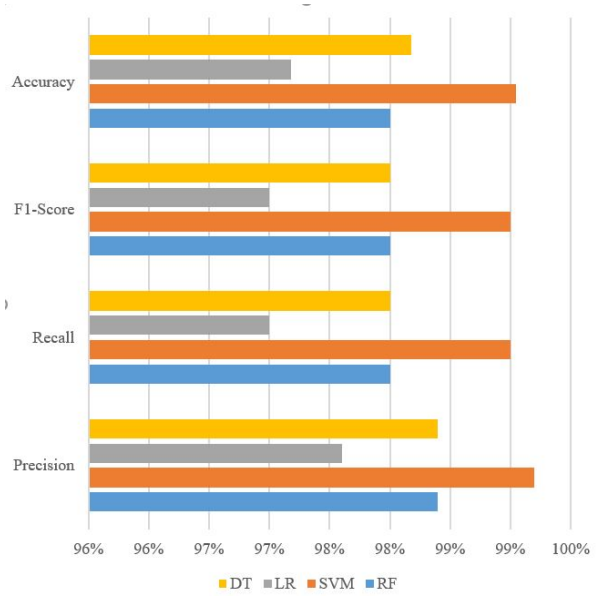


Fig. 10. Performance Measure of used algorithms

## V. CONCLUSION

This study proposes a high accuracy model for the detection of Acute Lymphoblastic Leukemia cancer and also analyzes different algorithms to compare performances. This model can achieve the desired accuracy to implement it in real life. The primary objective of this study was to implement a model to automatically detect leukemia cancer. Moreover, different algorithms were applied to identify the best suited approach to detect ALL cancer. After a fair comparison, Support Vector Machine (SVM) is recommended to classify blast cells as it provides higher precision and accuracy.

Image processing in medical sector along with machine learning/data mining approaches have been playing a key role in recent years for detecting various diseases automatically and intelligently. Experiments of such types as presented here cannot completely draw conclusion of detecting diseases, but help the medical practitioners to take their decisions appropriately. There are various other data mining approaches that could be applied to justify the results obtained so far. It is our future plan to apply evolutionary approaches from AI to conduct more experiments in order to get proper insight from the datasets.

## REFERENCES

[1] B. Ananya, A. Prabisha, and V. Kanjana, "Novel Approach to Find the Various Stages of Chronic Myeloid Leukemia Using Dynamic Short Distance Pattern Matching Algorithm," in *2018 3rd International Conference for Convergence in Technology (I2CT)*. IEEE, apr 2018, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8529812/

[2] R. Bagasjvara, I. Candradewi, S. Hartati, and A. Harjoko, "Automated detection and classification techniques of Acute leukemia using image processing: A review," in *2016 2nd International Conference on Science and Technology-Computer (ICST)*. IEEE, oct 2016, pp. 35–43. [Online]. Available: http://ieeexplore.ieee.org/document/7877344/

[3] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov 1986.

[4] N. Chatap and S. Shibu, "Analysis of blood samples for counting leukemia cells using support vector machine and nearest neighbour," *IOSR Journal of Computer Engineering*, vol. 16, pp. 79–87, 01 2014.

[5] C. Fatichah, M. L. Tangel, M. R. Widyanto, F. Dong, and K. Hirota, "Interest-Based Ordering for Fuzzy Morphology on White Blood Cell Image Segmentation," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 16, no. 1, pp. 76–86, jan 2012. [Online]. Available: https://www.fujipress.jp/jaciii/jc/jacii001600010076

[6] D. Foran, D. Comaniciu, P. Meer, and L. Goodell, "Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 4, pp. 265–273, 2000. [Online]. Available: http://ieeexplore.ieee.org/document/897058/

[7] M. Hahnle, F. Saxen, M. Hisung, U. Brunsmann, and K. Doll, "FPGA-Based Real-Time Pedestrian Detection on High-Resolution Images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2013, pp. 629–635. [Online]. Available: http://ieeexplore.ieee.org/document/6595939/

[8] P. Jagadev and H. Virani, "Detection of leukemia and its types using image processing and machine learning," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. IEEE, may 2017, pp. 522–526. [Online]. Available: http://ieeexplore.ieee.org/document/8300983/

[9] K. S. Kim, P. K. Kim, J. J. Song, and Y. C. Park, "Analyzing blood cell image to distinguish its abnormalities (poster session)," in *Proceedings of the eighth ACM international conference on Multimedia - MULTIMEDIA '00*. New York, New York, USA: ACM Press, 2000, pp. 395–397. [Online]. Available: http://portal.acm.org/citation.cfm?doid=354384.354543

[10] V. Kovalev, A. Grigoriev, and Hyo-Sok Ahn, "Robust recognition of white blood cell images," in *Proceedings of 13th International Conference on Pattern Recognition*. IEEE, 1996, pp. 371–375 vol.4. [Online]. Available: http://ieeexplore.ieee.org/document/547448/

[11] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in *2011 18th IEEE International Conference on Image Processing*. IEEE, sep 2011, pp. 2045–2048. [Online]. Available: http://ieeexplore.ieee.org/document/6115881/

[12] D. Majumder and M. Das, "An analytical approach for leukemia diagnosis from light microscopic images of Rbcs (Computational approach for leukemia diagnosis)," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, oct 2016, pp. 21–24. [Online]. Available: http://ieeexplore.ieee.org/document/7877383/

[13] S. Manago, C. Valente, P. Mirabelli, M. Napolitano, D. Corda, and A. De Luca, "Identification and classification of acute lymphoblastic leukemia cells from peripheral blood by using Raman spectroscopy," in *18th Italian National Conference on Photonic Technologies (Fotonica 2016)*. Institution of Engineering and Technology, 2016, pp. 78 (4 .)–78 (4 .). [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/cp.2016.0938

[14] A. M. Mehdi, M. S. Sehgal, A. Zayegh, R. Begg, and A. Manan, "K-Means Clustering on 3rd order polynomial based normalization of Acute Myeloid Leukemia (AML) and Acute Lymphocyte Leukemia (ALL)," in *2009 Third International Conference on Electrical Engineering*. IEEE, apr 2009, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/5173170/

[15] S. Mohapatra, D. Patra, S. Kumar, and S. Satpathy, "Lymphocyte image segmentation using functional link neural architecture for acute leukemia detection," *Biomedical Engineering Letters*, vol. 2, no. 2, pp. 100–110, jun 2012. [Online]. Available: http://link.springer.com/10.1007/s13534-012-0056-9

[16] S. Mohapatra, D. Patra, and S. Satpathy, "Automated leukemia detection in blood microscopic images using statistical texture analysis," in *Proceedings of the 2011 International Conference on*

*Communication, Computing & Security - ICCCS '11*. New York, New York, USA: ACM Press, 2011, p. 184. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1947940.1947980

[17] N. Patel and A. Mishra, "Automated leukaemia detection using microscopic images," *Procedia Computer Science*, vol. 58, pp. 635 – 642, 2015, second International Symposium on Computer Vision and the Internet (VisionNet'15). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915021936

[18] V. Piuri and F. Scotti, "Morphological classification of blood leucocytes by microscope images," in *2004 IEEE International Conference onComputational Intelligence for Measurement Systems and Applications, 2004. CIMSA*. IEEE, 2004, pp. 103–108. [Online]. Available: http://ieeexplore.ieee.org/document/1397242/

[19] B. Prasad, J.-s. Choi, and W. Badawy, "A High Throughput Screening Algorithm for Leukemia Cells," in *2006 Canadian Conference on Electrical and Computer Engineering*. IEEE, 2006, pp. 2094–2097. [Online]. Available: http://ieeexplore.ieee.org/document/4055027/

[20] J. Rawat, H. S. Bhadauria, A. Singh, and J. Virmani, "Review of leukocyte classification techniques for microscopic blood images," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, March 2015, pp. 1948–1954.

[21] O. Selcuk and F. Ozen, "Acute lymphoblastic leukemia diagnosis using image processing techniques," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, may 2015, pp. 803–806. [Online]. Available: http://ieeexplore.ieee.org/document/7129950/

[22] V. Shankar, M. M. Deshpande, N. Chaitra, and S. Aditi, "Automatic detection of acute lymphoblasitc leukemia using image processing," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE, oct 2016, pp. 186–189. [Online]. Available: http://ieeexplore.ieee.org/document/7887948/

[23] Sung-Huai Hsieh, Zhenyu Wang, Po-Hsun Cheng, I-Shun Lee, Sheau-Ling Hsieh, and Feipei Lai, "Leukemia cancer classification based on Support Vector Machine," in *2010 8th IEEE International Conference on Industrial Informatics*. IEEE, jul 2010, pp. 819–824. [Online]. Available: http://ieeexplore.ieee.org/document/5549638/