2018 Sixteenth International Conference on ICT and Knowledge Engineering

Leukemia Prediction from Microscopic Images of Human Blood Cell Using HOG Feature Descriptor and Logistic Regression

Hossain Abedy[†], Faysal Ahmed[‡], Md. Nuruddin Qaisar Bhuiyan[▲],
Maheen Islam[▼], Md. Nawab Yousuf Ali[※], and Md. Shamsujjoha[*]

[†‡▲▼※*]Department of Computer Science and Engineering, East West University, Dhaka-1212, Bangladesh.

Email: [†]abedy.ewu@gmail.com, [‡]faysalahmedewu108@gmail.com, [▲]qaisar.ewu.cse14@gmail.com,
[▼]maheen28_i@yahoo.com, [※] nawab@ewubd.edu, and [*]dishacse@yahoo.com

***Abstract*—Leukemia originates in bone marrow. It massively affects the production of appropriate blood cells. Hence, its early detection is very crucial for human living. Generally, computational approaches for Leukemia detection use microscopic blood cells images. Then, machine learning based models are trained and tested for accurate measurement. The main challenge here is to achieve an acceptable accuracy with a scalable method. However, data inconsistency, missing values and data incompleteness made the researchers' job much more difficult. In these consequences, this paper proposes a scalable Leukemia prediction method based on a publicly available ALL_IDB dataset using the HOG feature descriptor and Logistic Regression. Initially, the proposed method used Canny edge detector and noise reduction operators to detect the exact shape of Lymphocytes. Then, Principal Component Analysis (PCA) is applied to the detected image shapes. The PCA reduces the data dimensions without losing any valuable information and thus greatly minimizes the afterward computational cost. Finally, a classifier based model is produced for unforeseen events and it is tested. The results are validated using *n*-fold cross-validation technique, where *n* is a positive integer greater than or equal to three. The maximum average accuracy of the proposed model is 96% which is much higher than the state-of-the-art schemes.**

***Keywords*— HOG Feature Descriptor, Image Processing, Leukemia, Logistic Regression, Machine Learning, Prediction**

I. INTRODUCTION

Acute Lymphocytic (or Lymphoblastic) Leukemia is the result of overproduction and multiplication of immature lymphocytes within white blood cells. There are two types of Acute Leukemia, (i) Acute Lymphoblastic Leukemia (ALL) and (ii) Acute Myeloid Leukemia (AML). Both Leukemia originate in bone marrow causing rapid growth of immature Lymphocytes. This thoroughly hampers the production of red blood cell, white blood cells and platelets in the human body[1]. Generally, Leukemia spread rapidly into the bloodstream and can be life-threatening when left untreated. It has been shown that the early diagnosis of Leukemia can prevent its spreading and are very helpful for the patient's recovery especially in the case of children [2]. Symptoms of Leukemia are very much similar to the normal bacterial and virus-related disease such as fever, weakness, severe infections, weight loss, recurrent nosebleeds, and small red spots in the skin. Hence, its early detection and diagnosis are very difficult.

On one hand, human expert approaches of Leukemia detection consist of several observational stages but need individual supervision and are much more time-consuming. Moreover, the results of the expert approaches are subjective and imprecise as the whole process relies on the operator's skills. This might also be subjected to unwanted errors. On the other hand, computational methods for Leukemia detection automatically analyze the blood cells and its components from microscopic images. This analysis involves cell classification and blast counting. Generally, healthy or infected blood cell features are extracted from an image dataset based on the morphological analysis. Then, an algorithm is employed to predict the condition of a blood cell and classify it accordingly. However, feature extraction from the representative images is a challenging task for the researchers, where accuracy is a major concern. In addition, unavailability of the dataset, incomplete data, overlapping data and other components make it even harder to extract the required features from the image dataset. However, computational methods used segmentation, training and validation before testing a classifier model. Thus, it is expected that the automated scheme overcomes the drawbacks of the human method with better efficiency, accuracy and provide the final outcome in an acceptable time frame.

In these consequences, this paper proposes an automatic computational machine learning-based classifier model to predict Leukemia from microscopic images of human blood cell using the publicly available ALL_IDB dataset [3]. The proposed method treats the leukocytes of blood cells as objects and then retrieves these objects using HOG descriptor along with the edge detector. The purpose of using an edge detector is to identify the shape of the cell blasts which is then converted into HSV color format within a pre-defined range. Then, the dimension reduction algorithm is applied to project the features from higher to a lower dimension. Finally, the logistic regression is used to train and test the model for classification purpose. The rest of the paper is organized as follows: Sec. II briefly discusses the literature which is required to understand the effectiveness of the proposed work. Sec. III shows the design and working procedures of the proposed HOG feature descriptor and logistic regression based Leukemia prediction model. A brief performance study is also presented in this section for evaluation of the proposed work. Based on all these discussions, a conclusion is drawn in Sec. IV.

978-1-5386-7159-7/18/$31.00 ©2018 IEEE

II. LITERATURE REVIEW

This section discussed the necessary background which is required to understand the proposed work. Initially, the section presents discussed about the dataset ALL-IDB that has been used for the proposed model. Then, a brief survey on computational methods for Leukemia prediction is presented.

A. The Dataset

The proposed research use ALL-IDB dataset. It consists of a microscopic image of human blood samples. It is publicly available and free for the evaluation and the comparison for newly developed algorithms. This dataset has massively been used by the researchers for segmentation and image classification. As mentioned by the authors [3], images of All-IDB dataset are classified (positioned) by the expert oncologists. The images of ALL-IDB dataset are captured using an optical microscope, having 24-bit color depth and resolution is 2592×1944.

B. Survey on Existing Approaches for Leukemia Prediction

In literature, a good number of works have put efforts to detect Leukemia from the microscopic images. In these works, a low-pass filter is used to remove noise from the background and then white blood cells are segmented with different threshold operations and image clustering. Piuri *et. al.*, proposed an approach based on detecting edges for white blood cells segmentation [1]. A threshold selection method is shown for leukemia detection in [4] by Otsu, whereas Cseke proposed a fast segmentation scheme with automatic threshold selection by a recursive procedure in [5]. The recursive procedure is defined by maximizing the interclass variance between dark, gray, and bright regions of an image. Pattern recognition methods based on multispectral imaging is shown in [6]. This method segment images uses Support Vector Machine (SVM). Here, SVM is applied directly to the spectrum of each pixel and using sequential minimal optimization algorithm for feature selection to reduce the time of training SVM classifier. The author claims that the method is robust but can't handle data uncertainty. A survey on automated microscopic image analysis for leukocytes identification is shown in [7]. Here, authors categorized, evaluated, and discussed most prominent works for automatic Leukemia prediction till 2014 and then developed methods for leukocyte identification. A future research perspective had also been presented in [7]. Segmentation and statistical texture analysis based Leukemia detection are shown in [8] and [9], respectively. In [10] and [11], K-means clustering and Fuzzy C means based Leukemia prediction is discussed. In [10], a color adjustment step was introduced before the image segmentation 300 microscopic blood smear images. Then, color space decomposition and k-means clustering were combined for segmentation. Ko *et. al.*, introduces stepwise merging rule on clustering and boundary removal rules with a gradient vector flow (GVF) snake for the segmentation of white blood cells in [12] for automatic Leukemia identification. Transformed color space such as HSV and HIS were introduced by researchers for Leukemia prediction in [13-15]. More discussions on the existing approaches of automatic Leukemia predictions are beyond the scope of this paper. The interested reader can see [2, 7, 9] for supervised and [8, 10] for unsupervised approaches.

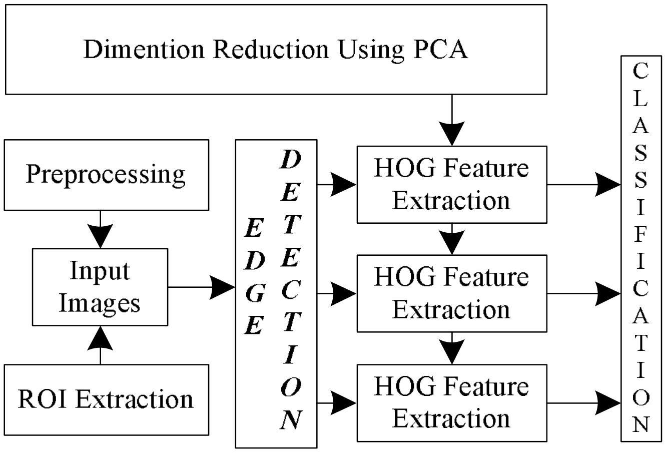

Figure 1. An architectural block diagram of the proposed model

III. PROPOSED WORK

An architectural block diagram of the proposed HOG feature descriptor and logistic regression based automatic Leukemia prediction model is shown in Fig 1. The working procedure of this model is as follows:

Stage 1, Preprocessing and region of interest (ROI) extraction: ALL-IDB dataset contains images of different height and width. Initially, the proposed model resized these images into 1000×1000 pixels. Then the resized images are converted into HSV color space and filter the purple color. This color is the ROI. At the end of this stage, the images are converted to a grayscale image. In Figs. 2, 4 sample blasts and non-blast cell images are shown along with its corresponding ROI and HSV color format.

Stage 2, Canny edge detection: The proposed model applies the Canny edge detection algorithm with gaussian filter to reduce noise. It also uses a gradient operator like Sobel and computes the threshold gradient to detect edges. The produces images of non-blast and blast cell edges are shown in Fig. 3.

Stage 3, HOG feature extraction: In this stage, the histogram of Oriented Gradients (HOG) feature is calculated from the output images of the previous stage. The execution result of the proposed method for ALL-IDB dataset is shown in Fig. 5. The dimension achieved for the HOG feature vector during the execution of the proposed method is 3,52,836.

Stage 4, Dimension reduction by PCA: The previous stage produces images with huge dimension and hence is very difficult and time-consuming to work with these images. Thus, the proposed method reduced the dimension if the output images of Stage 3 using PCA with an explained variance ratio of 0.95. The reduced dimensions of feature vectors for five different executions are listed in Table I.

Stage 5, Classification by Logistic Regression: Finally, the proposed model is trained using the logistic regression classifier. The proposed model is trained and tested based 3-fold cross validation. Corresponding confusion matrix is shown in Table II. From these tables we find that the proposed method is a scalable and reliable computational approach for blood cancer (Leukemia) prediction.

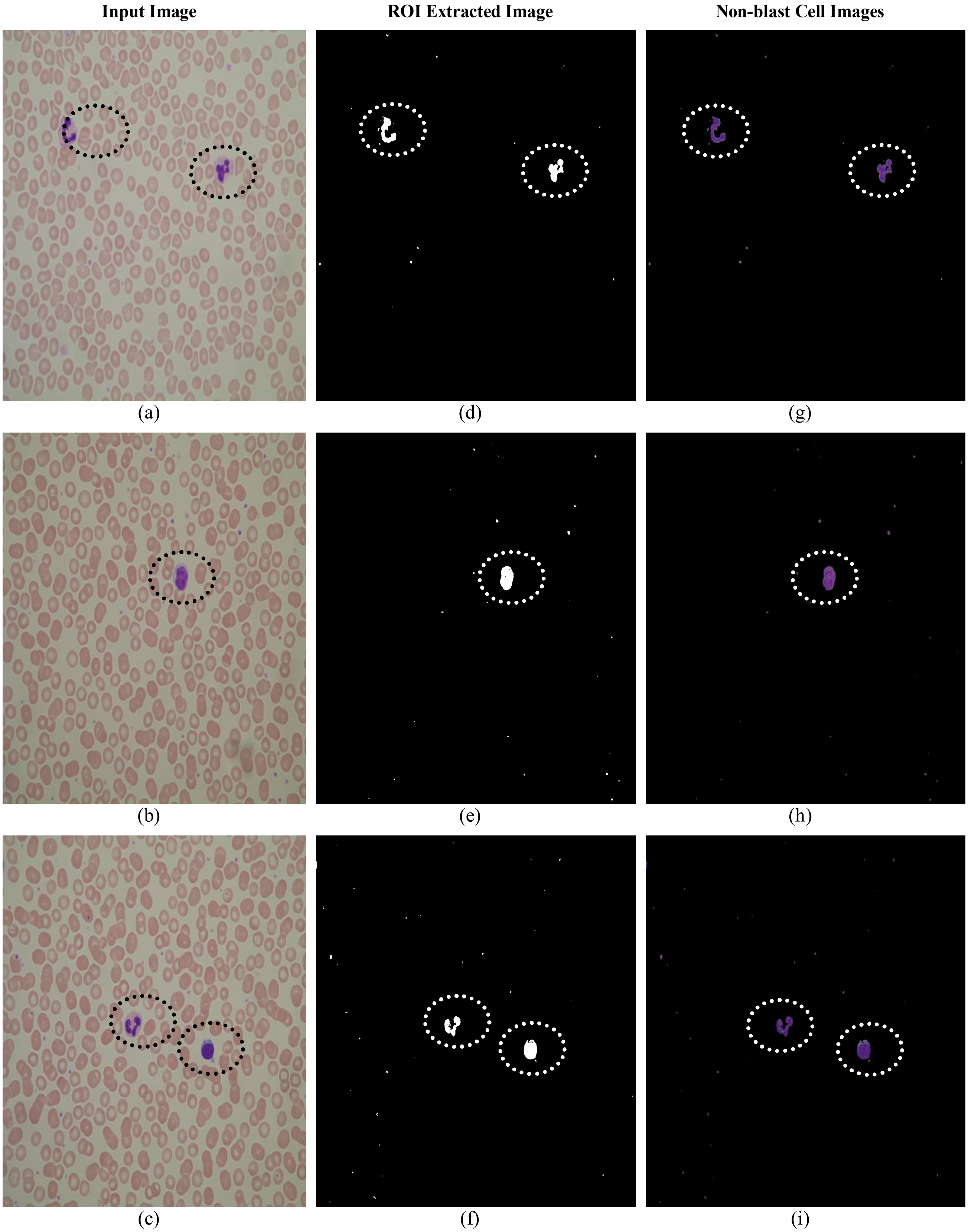

Figure 2. Stage 1 of the proposed model (a) (b) (c) Input images, (d) (e) (f) Corresponding ROI extracted images of (g) (h) (i) Corresponding non-blast cell images of (a) (b) (c) respectively

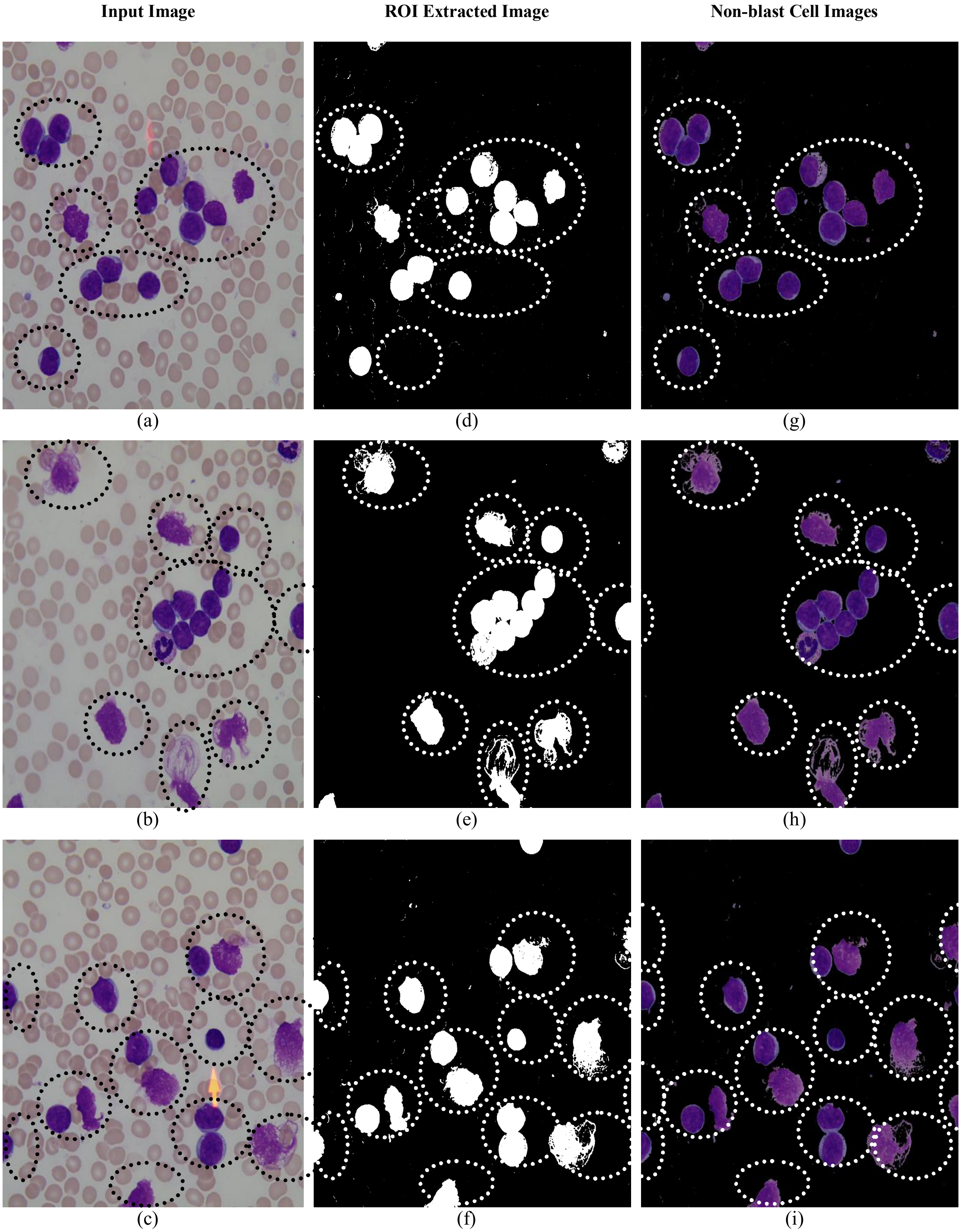

Figure 3. Stage 1 of the proposed model (a) (b) (c) Input images, (d) (e) (f) Corresponding ROI extracted images of (g) (h) (i) Corresponding non-blast cell images of (a) (b) (c) respectively

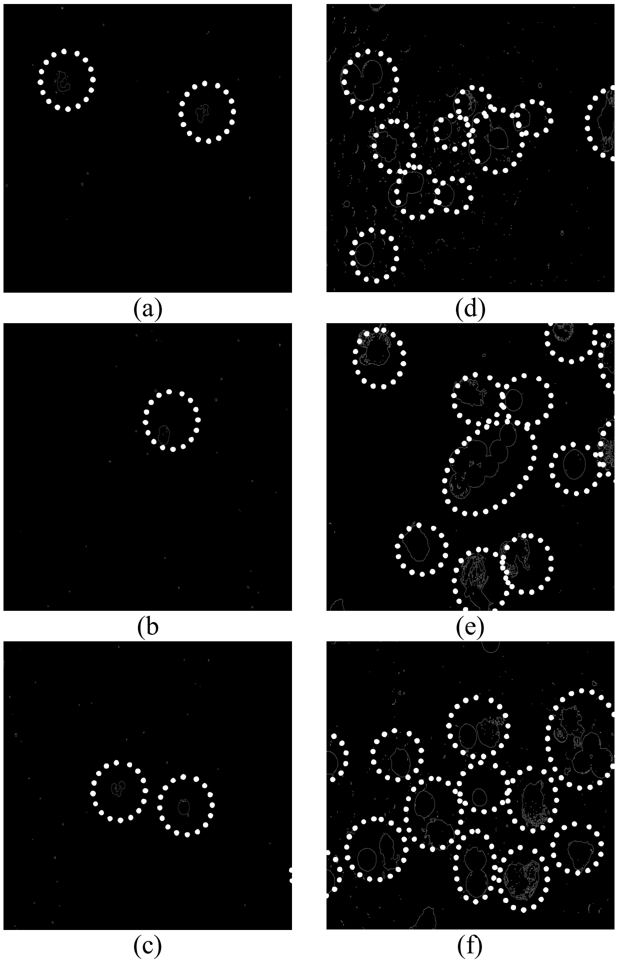

Figure 4. Canny edge detection (a)(b)(c) non-blast cell,(d)(e)(f) blast cell

TABLE I. PERFORMANCE OF THE PROPOSED MODEL (FEATURE DIMENSION 51 AND 52) FIVE DIFFERENT EXECUTION

Validation Accuracy	**Test Accuracy**	**Precision**	**F1-Score**
94	91	81.82	90
95	100	100	100
97	91	80	88.89
94	95	91.67	95.65
98.8	100	100	100

TABLE II. CONFUSION MATRIX FOR THE DATA OF TABLE I (FIVE DIFFERENT EXECUTION)

True Negative	**False Positive**	**False Negative**	**True Positive**
9	2	0	11
10	0	0	12
8	2	0	12
11	1	0	10
12	0	0	10

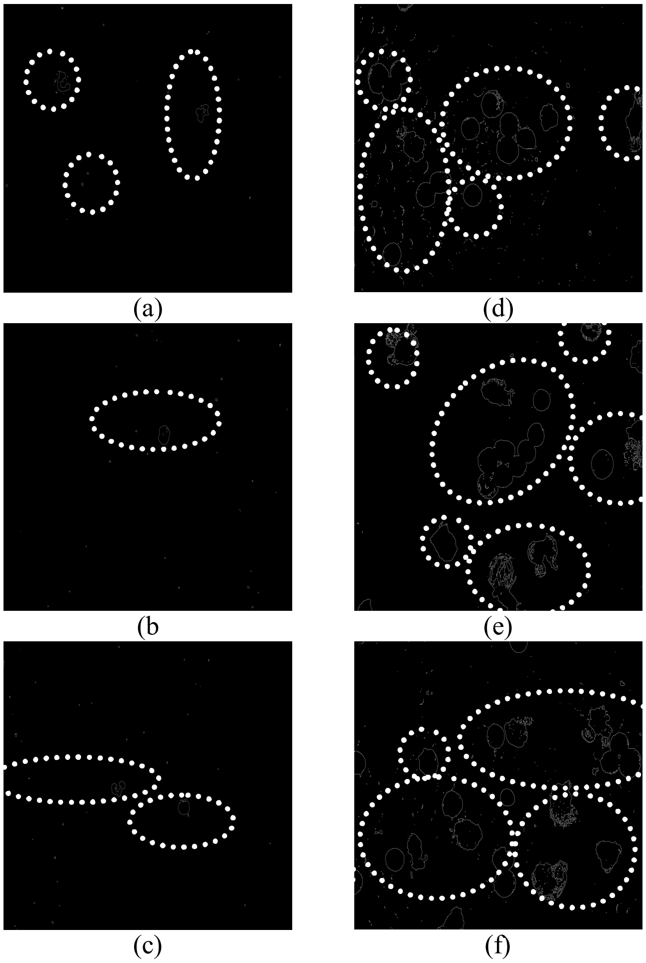

Figure 5. HOG Images (a)(b)(c) non-blast cell,(d)(e)(f) blast cell

IV. CONCLUSION

This research used microscopic images from ALL_IDB dataset to classify the blast and non-blast cells for Leukemia prediction. The proposed model initially detects the shape of the blast cell from images. Then, the Gaussian filter is used to remove noise, Sobel kernel for image filtering, PCA for data dimension reduction of the feature vector etc. Finally, the HOG feature descriptor and logistic regression is used for the classification. The performance of the proposed method is shown in Tables I and II, which proves its significance of the computational research works to predict Leukemia. An interesting future work can be evaluating the performance until stage 4 of the proposed model based on Artificial Neural Network, Bayesian Network and other Machine Learning based algorithm for classification.

REFERENCES

[1] V. Piuri and F. Scotti, “Morphological classification of blood leucocytes by microscope images”, 2004 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2004. CIMSA..

[2] F. Scotti, “Automatic morphologic analysis for acute leukemia identification in peripheral blood microscope images”,2005 IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications, 2205. CIMSA..

[3] ALL-IDB, Acute Lymphoblastic Leukemia Image Database for Image Processing, Available at "https://homes.di.unimi.it/scotti/all/" Last Accessed on March 2 2018

[4] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979.

[5] I. Cseke, "A fast segmentation scheme for white blood cell images", Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. IV. Conference D: Architectures for Vision and Pattern Recognition,.

[6] N. Guo, L. Zeng and Q. Wu, "A method based on multispectral imaging technique for White Blood Cell segmentation", Computers in Biology and Medicine, vol. 37, no. 1, pp. 70-76, 2007.

[7] M. Saraswat and K. Arya, "Automated microscopic image analysis for leukocytes identification: A survey", Micron, vol. 65, pp. 20-33, 2014.

[8] C. Zhang, X. Xiao, X. Li, Y. Chen, W. Zhen, J. Chang, C. Zheng and Z. Liu, "White Blood Cell Segmentation by Color-Space-Based K-Means Clustering", Sensors, vol. 14, no. 9, pp. 16128-16147, 2014..

[9] S. Mohapatra, D. Patra and S. Satpathy, "Automated leukemia detection in blood microscopic images using statistical texture analysis", Proceedings of the 2011 International Conference on Communication, Computing & Security - ICCCS '11, 2011.

[10] N. Salem, "Segmentation of white blood cells from microscopic images using K-means clustering", 2014 31st National Radio Science Conference (NRSC), 2014.

[11] Puttamadegowda J. and Prasannakumar S. C., "White Blood cell sementation using Fuzzy C means and snake", 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016.

[12] B. Ko, J. Gim and J. Nam, "Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake", Micron, vol. 42, no. 7, pp. 695-705, 2011.

[13] K. A.ElDahshan, M. I. Youssef, E. H. Masameer and M. A. Mustafa, "Segmentation Framework on Digital Microscope Images for Acute Lymphoblastic Leukemia Diagnosis based on HSV Color Space", International Journal of Computer Applications, vol. 90, no. 7, pp. 48-51, 2014.

[14] K. ElDahshan, M. Youssef, E. Masameer and M. Hassan, "Comparison of Segmentation Framework on Digital Microscope Images for Acute Lymphoblastic Leukemia Diagnosis Using RGB and HSV Color Spaces", Journal of Biomedical Engineering and Medical Imaging, vol. 2, no. 2, 2015.

[15] V. Singhal and P. Singh, "Correlation based Feature Selection for Diagnosis of Acute Lymphoblastic Leukemia", Proceedings of the Third International Symposium on Women in Computing and Informatics - WCI '15, 2015.