Detection of Leukemia Using K-Means Clustering and Machine Learning

V. Lakshmi Thanmayi A(✉), Sunku Dharahas Reddy, and Sreeja Kochuvila

Department of Electronics and Communication Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, Bengaluru, India
avlt484@gmail.com, k_sreeja@blr.amrita.edu

Abstract. Leukemia or blood cancer is a common and serious disease across countries which is caused due to the sudden increase in White Blood Cells (WBCs) in blood. This increase in WBC is due to the production of immature or blast cells in the bone marrow of the affected person. Detection and diagnosis at early stage is important. Additionally, computer-aided diagnosis will enhance the process of detection with better accuracy. In this paper, we developed an algorithm for early-stage detection of leukemia using image processing. We also used machine learning classification techniques to classify between cancerous and non-cancerous cells. The algorithm uses K-means clustering for the segmentation of images and a linear Support Vector Machine (SVM) classifier for the classification. ALL-IDB data set has been used to validate the algorithm. A total of 368 images are used in the algorithm. Algorithm offers 95% of accuracy and an approximate 93% of precision.

Keywords: Leukemia · White Blood Cells · Machine learning · Support Vector Machine

1 Introduction

Cancer [1], the most common disease across the world, is caused due to the rapid multiplication of abnormal cells in the body. This reduces the functioning of the body. Cancer is of many types based on the cells attacked. Leukemia is a type of cancer which affects the white blood cells in the blood and thus called blood cancer [2]. Leukemia affects the bone marrow, the place of production of blood cells, causing an imbalance in the blood cell count. These increased abnormal cells lack the ability to fight against infection and affect the way healthy organs work.

Leukemia, if treated at the early stage, can stop the abnormal cells from increasing rapidly. Treatment at an early stage calls for the diagnosis of the disease at an early stage. The worldwide accepted methods in the diagnosis of leukemia include the physical examinations and lab tests [3]. During the physical examination, doctors look for swollen or bleeding gums and tiny rashes, which

N. Kumar et al. (Eds.): UBICNET 2021, LNICST 383, pp. 198–209, 2021.
https://doi.org/10.1007/978-3-030-79276-3_15

can be common with illnesses like flu. In the labs, the blood samples are analyzed by experts under powerful microscopes where the total blood count, including red blood cells, white blood cells, and platelets, are obtained. They also physically analyze the cells for the release of any substance that shows the presence of cancer. The diagnosis in the above methods can be effective only when the actual symptoms of the disease are seen. An algorithm for the detection of the disease even before any symptoms are visible is required. Many image processing algorithms have been developed to prove the presence of cancer at an early stage.

Digital image processing has a vast application in the diagnosis in the medical field. These image processing techniques have numerous advantages, for example, data flexibility, versatility, and estimation, information storing, and correspondence. Some significant machine learning classifiers include [4,5]: Linear classifiers, tree based classifiers, Support Vector Machine (SVM), Gaussian naive Bayes classifiers and Stochastic Gradient Descent (SGD) classifiers. SVM is a supervised learning strategy that can be utilized for both classification and regression. SVM accepts the input and output data points to create a hyperplane, also called a decision boundary which differentiates the input data based on the output classes. SVM can be used for both binary and multi-class classification.

In [6], a segmentation of nuclei cells and their classification algorithm is explained. Of the many segmentation techniques, the paper highlights the usage of K-means clustering method as a color based segmentation method, for image segmentation. The pixels are classified based on the *a and *b component of the L*a*b (luminosity, chromaticity layer-a, and chromaticity layer-b) color space. The paper features the significance of Hausdorff Dimensions during feature extraction. It helps in calculating roughness, perimeter and other parameters thus helping in the classification of images. Authors in [7] explains the classification of different types of cancer, Acute Myeloid Leukemia (AML) and Acute Lymphocytic Leukemia (ALL) are analysed depending on how leukemic cells look under the magnifying lens and the kind of cell included. While Chronic Myeloid Leukemia (CML) and Chronic Lymphocytic Leukemia (CLL) are analysed depending on the WBCs tally at the hour of conclusion. The presence of immature WBCs, or myoblasts in the blood and bone marrow is also used to conclude for AML and CML. This paper uses a mathematical operations based segmentation like addition and subtraction of various pre-processing of an image. The feature extraction of the segmented image is done using a MATLAB function "regionprops". Classification of the features is obtained using a SVM classifier for efficient results.

In this paper, we present an algorithm which gives an initial conclusion of the possible presence of cancer by a computerized analysis of digital, microscopic blood images. Image pre processing filters have been used to enhance the quality of the image for an output with lesser error. After the image is preprocessed, segmentation algorithms like: edge detection, watershed transform, K-means clustering, thresholding, are used to separate out the required part of the image from the whole image. We use the K-means clustering algorithm for

the nuclei segmentation. Then, the features are extracted from the segmented images. These extracted features are used for classification in a machine learning algorithm. We show the binary classification of the data using a SVM classifier. The extracted features are considered as an input data and this input data is split into train and test sets which are used in classification of the images as cancerous and non-cancerous. The existing methods consider more number of features during the feature extraction stage for classification. In our proposed work we extract minimum number of features and analyse them to attain a greater accuracy. This reduces the computational power while extracting the features.

This paper is organized as follows. Section 2 gives few recent and relevant work in the area of leukemia detection. Section 3 describes the data sets and the steps to process the images. The proposed algorithm is explained in Sect. 4 followed by classification technique in Sect. 5. Results are analysed and discussed in Sect. 6. The conclusion and future scope is given in Sect. 7.

2 Few Recent and Relevant Work

In this section, we discuss few recent and very relevant work related to our work. A nucleus extraction method from the cytoplasm of the WBC as in [8] uses color conversion, intensity threshold and gradient method. An algorithm with color segmentation and Otsu's segmentation techniques for the separation process has been used. A computer aided diagnosis to recognize ALL kind of leukemia is created in [9]. In this, authors have discussed the proper feature extraction methods from the core of the WBCs for examination purposes. The paper deals with the appropriate feature extraction techniques from the nucleus of the WBCs for analysis purpose. Feature extraction is obtained using discrete cosine transform which helps in dividing the images into parts.

Authors in [10] used K-means clustering to extract the regions of interest along with basic enhancement, morphology filtering and segmentation technique. An adaptive histogram equalization is used for image pre-processing. The performance parameters in this paper have been analyzed using probability random index, this parameter gives us the exactness of segmentation.The paper explains in detail the various features in feature extraction. It explains the different set of values one gets during estimation which is helpful in the classification of the images. In [11] image pre-processing using median filters, conversion from Red Green Blue (RGB) to Hue, Saturation and Value (HSV) and thresholding have been performed. The boundaries of different image areas are found so that location, features and shape can be found in the image using the integral projection algorithm used in feature extraction. The paper uses watershed technique which is characterised by mountains and valleys. The mountains represent high intensity and valleys represent low intensity.

In [12], arithmetic operations and image enhancement techniques are used for segmentation of the nucleus from white blood cells. Also the problem of thresholding and K-means clustering can be solved by using this method. The

image intensity is adjusted by using linear contrast stretching which segments out the nucleus. The blast cells from the normal lymphocyte cells are classified using a KNN classifier. In [13], the diagnosis of ALL type of cancer is explained by converting the RGB image into Cyan, Magenta, Yellow and Black (CMYK) scale as part of image pre-processing. Zack's algorithm, a triangle oriented threshold method, is used to segment out the WBCs. In this method a straight line is generated between the maximum and minimum value of the image histogram. After this an ideal threshold is determined and separation is completed utilizing the obtained threshold values.

In [14], important steps such as pre-processing, segmentation and match-making have been used. Otsu's image segmentation is used for segmenting the leukemia smear image database and matching of image pattern is done using Maximally Stable Extremal Regions (MSER). Pre-processing in this paper is completed using selective median filtering with unsharp masking and contrast enhancement techniques. MSER although having limited performance over blurred and/or textured images it has advantages over methods like moderate computational complexity, and it is more suitable for hardware implementation due to its algorithmic structure. Otsu's method maximises the between class variance by having an exhaustive search to evaluate this criteria. Segmentation in [15] is performed with morphological operators and Otsu's thresholding. Then, utilization of nucleus features with supervised KNN classifiers is used for the classification of the extracted features. In this process, the RGB images are converted into gray-scale to reduce the computational complications. Linear contrast enhancement and histogram equalization have been used for increasing the image intensity such that the total image, except for the nucleus, is brightened. Morphological erosion and closing operations to better the performance of segmentation and feature extraction processes is performed.

A fast correlation based filter has been used in [16] to select the most prominent gene for the feature extraction. This method is used to reduce the huge data set for classification. A SVM classifier is then implemented to classify the tumour cells. In [17] authors find different stages of CML using dynamic short distance pattern matching algorithm using the normal and abnormal gene sequences.

3 Typical Image Processing and Input Database

3.1 Typical Image Processing

The algorithm for the detection of leukemia normally involves multiple stages of image processing techniques which is shown in Fig. 1. These steps are used in sequence on the data set using MATLAB to obtain the best result. The conclusion of leukemia, either through the lab and manual tests or through image processing techniques, is obtained on microscopic blood smears. These blood smears are examined to check for any variations or abnormalities from normal blood cells. During blood tests a sample from the patient's blood is seen under a powerful magnifying lens or microscopes which amplify the blood smear as per the prerequisite. These microscopic blood samples are digitized into

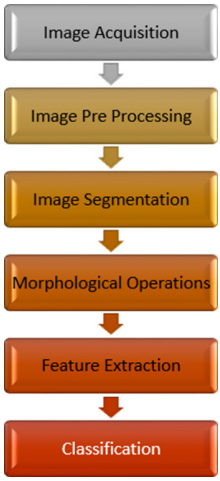


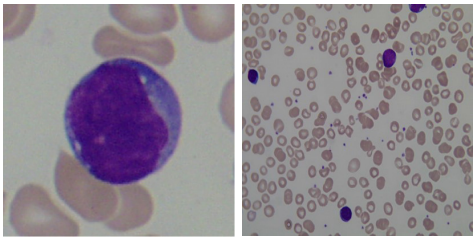
Fig. 1. Typical steps in image processing and classification

pictures with the goal that the image processing can be performed without any problem. These images of blood smears are accessible legitimately from the lab for a specific foundation purpose or are accessible online in picture banks where individuals can download for research purposes.

Image pre-processing is a way towards improving the image data, suppressing the undesired contortions, and enhancing the image features so that the analysis of the image at further stages of the algorithm yields a better error-free result. The images are made free of noise, de-blurred and other refinement processes are done as required by the procedure decided on an application use. Segmentation is a significant phase of the image processing process, since it removes the unwanted objects, separating the objects of our requirement, for additional processing. The major practical application of segmentation lies in the classification of the pixels where each pixel is assigned some labels. Pixels with similar labels share common characteristics.

Morphological changes are some operations dependent on the picture shape. It is generally performed on binary images. It needs two information sources, the image on which morphological operations are performed and the structuring element or kernel which tells the type of operation to be performed. The two basic morphological operations are erosion and dilation. These two major operations have variations such as opening, closing, gradient and so on.

A picture has an immense informational collection that must be portrayed. The information in the picture is ordered into a lot of features which is called feature extraction. Feature extraction in machine learning, pattern recognition and in image processing, starts from initial data values obtained during the process and derives a new set of values which are more informative and non repetitive, helping in research, and classification which provide better human



(a) (b)

Fig. 2. Microscopic blood sample images (a) ALL-IDB1 (b) ALL-IDB2

conclusions. In this paper, the morphological features, related to the shape of the image, are analysed which helps in the classification in further steps.

After the features are extracted, the images or the data are classified into different classes or categories by analysing the features. An inter relation is established between these features through machine learning algorithms in order to classify the data easily and efficiently.

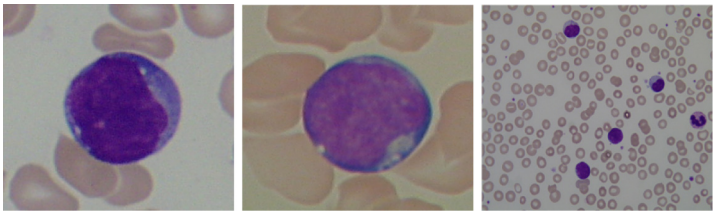
3.2 Input Database Images (IDB)

An online data set [18], ALL-IDB data set by Fabio Scotti from Università Degli Studi di Milano, has been used for leukemia detection in this paper. This data set includes ALL-IDB1 with 108 images. The nuclei from ALL-IDB1 have been separated in the form of images to form ALL-IDB2 which has 260 images. The algorithm has been implemented on both the data sets acquired. The example images of ALL-IDB1 and ALL-IDB2 are shown in Fig. 2.

4 Proposed Algorithm

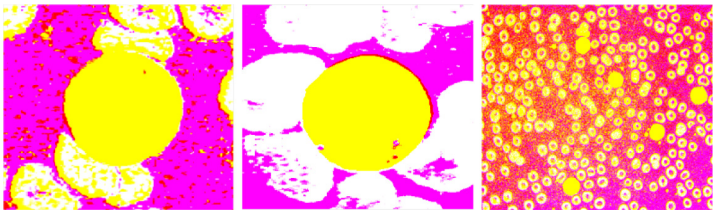
This paper implements blind de-convolution and Gaussian noise removal methods for the refinement of the images. The acquired images are in an RGB space and are converted into a different color space such as L*a*b, CMYK, HSV for image segmentation. The RGB space is connected with the measure of light hitting the item and the genuine contrasts are not appropriately noticeable. This paper converts the image in RGB space to L*a*b space for the process of segmentation. The input and the converted image into L*a*b space is shown in the Fig. 3 and Fig. 4 respectively.

The images after pre-processing are segmented using K-means clustering algorithm to separate out the nucleus from the cytoplasm of the WBCs. K-means clustering is an iterative, color based segmentation method which is used to part a picture into K number of groups or clusters. K centroids are initialised first and every one of the pixels is mapped into its closest centroid esteem. In the wake of grouping all the pixels, a new centroid for each cluster is formed. This method uses Euclidean distance to calculate the distance between the pixels and



(a) (b) (c)

Fig. 3. Image samples used for pre-processing (a) Sample-001 (b) Sample-090 (c) Sample-022



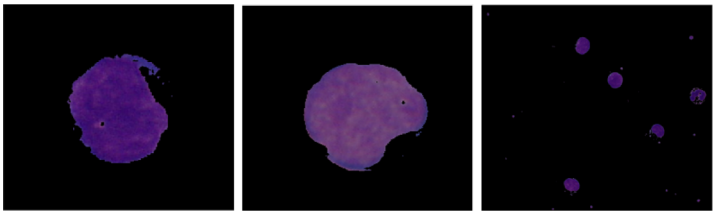
(a) (b) (c)

Fig. 4. Converted images from RGB to L*a*b (a) Sample-001 (b) Sample-090 (c) Sample-022

centroids to assign a pixel to a cluster with minimum distance. After the image is converted from RGB space, the colors are classified into K clusters. The process of clustering is carried out multiple times so as to avoid local minima. In this paper, the clustering process is performed three times where each component is separated out in each of the clustering processes. The nucleus component of the image is extracted after the clustering process is completed based on which cluster separates the nucleus best (as shown in Fig. 5).

After the process of segmentation, morphological operations are performed on the binary version of the segmented nucleus to make the image better analysable during feature extraction. Operations like closing, dilation, and opening of the image have been performed. Dilation is the process of adding pixels at the edges of the images in order to increase the white region of the images removing the shrunken look of the image. Opening is used to remove the minute noise, which is in the form of small white pixels over the black background, from the image. Closing is done when the small holes are there within the segmented image which are formed due to the variations in the color intensities. These holes are filled using the closing operation to form a solid, final image ready for feature extraction as shown in Fig. 6.

The segmented nucleus is analysed to extract the features. These features include energy, entropy, area, solidity and so on. In this paper we have analysed the nucleus to get the following features: area, perimeter, eccentricity, circularity,



(a) (b) (c)

Fig. 5. Segmented blue nuclei (a) Sample-001 (b) Sample-090 (c) Sample-022 (Color figure online)

(a) (b) (c)

Fig. 6. Morphologically processed nuclei (a) Sample-001 (b) Sample-090 (c) Sample-022

and solidity. These features have been calculated for the images in ALL-IDB1 and ALL-IDB2 and the same were tabulated. Machine learning algorithms are implemented on these data values for the classification of the images.

5 Classification

A supervised machine learning model SVM has related calculations that is used for classification and regression. Given a set of input data, where each of the set belongs to anyone of the output classes, the SVM classifier trains a portion of input data to create a model which assigns the new values to one of the output classes. These points are represented in space such that they are divided by a hyper-plane also called decision boundary. The hyper plane creates a boundary between classes. Two classes as in this case, so that the data points lie on either side of the plane.

The kernel of the classifier that is used is defined by the user. Kernels represent the type of SVM classifier to be used. Kernels in SVM include linear, nonlinear, polynomial, radial basis function, and sigmoid. Gamma parameter is generally used in a non-linear SVM classifier where its value affects the Gaussian variance. A 'C' parameter is defined in an SVM classifier which regulates the error produced during training and testing. This is changed in a way to get minimum error in both the cases.

The classification of data in this paper has been done using a linear SVM classifier. The features extracted from the segmented images are in the form of integer values. The combination of each feature set representing the image as one of the two classes cancerous and non-cancerous. These values form the input data for the classifier which are trained based on the output class values where cancerous is considered as 1 and non cancerous is considered as 0. The test data is analysed and is assigned to any one of the classes and a decision boundary is formed. The confusion matrix is found for validating the algorithm. A confusion matrix is a tabular representation of the classifier which consists of True Positive (TF), True Negative (TN), False Positive (FP), False Negative (FN). The accuracy, precision, specificity and recall is calculated from the confusion matrix using the following formulae.

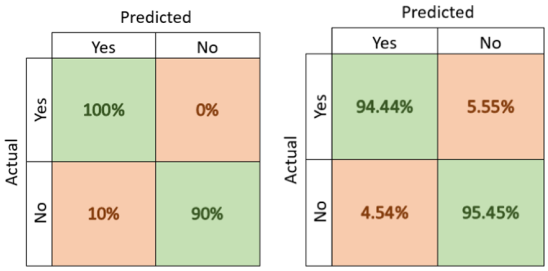
$$Accuracy = \frac{TP + TN}{Total}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Specificity = \frac{TN}{TN + FP}$$
$$Recall = \frac{TP}{TP + FN}$$

6 Results and Discussion

The image processing steps are carried out on an ALL-IDB data set as explained in Fig. 1. The ALL-IDB1 and ALL-IDB2 images as in Fig. 3 have been pre processed to generate images as shown in Fig. 4. These images have been segmented to separate out the nucleus using K-means clustering process Fig. 5. The segmented nucleus is morphologically processed to improve the pixel rate and the final image in Fig. 6 is analysed for feature extraction using MATLAB. A linear SVM classifier has been implemented in order to classify the images to compute accuracy and validate the algorithm.

180 images from ALL-IDB1 and 260 images from ALL-IDB2 have been analysed. 80% of these have been used for training and 20% for testing purposes. Figure 7(a) and Fig. 7(b) represent the confusion matrix for ALL-IDB1 and ALL-IDB2 respectively. The accuracy, precision, specificity and recall were calculated from the confusion matrix as shown in Table 1. Accuracy determines how often the classifier is correct. Precision and recall refer to the percentage of the results which are relevant and the percentage of total relevant results correctly classified by the algorithm respectively.

Accuracy gives the overall performance of the classifier. Hence different classifiers were used to calculate the accuracy and the results are compared in Table 2. A highest accuracy of 95.45% is obtained for ALL-IDB1 and 95% accuracy for ALL-IDB2 using the proposed algorithm.



(a) **(b)**

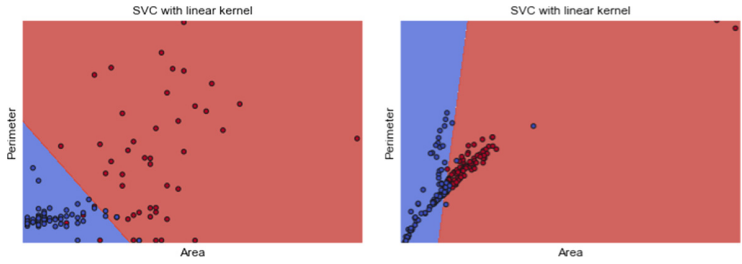
Fig. 7. Confusion matrices (a) ALL-IDB1 (b) ALL-IDB2

Table 1. Summary of different parameters.

Metrics	ALL-IDB1	ALL-IDB2
Accuracy	95.45%	95%
Precision	92.30%	94.44%
Specificity	90%	95.45%
Recall	100%	94.44%

Table 2. Comparison of accuracy using different classifiers

Classifier	Accuracy for ALL-IDB1	Accuracy for ALL-IDB2
SVM	95.45%	95%
KNN	81.8%	87.5%
Decision tree	90.69%	92.5%
Ada boost	90.9%	93.3%



(a) **(b)**

Fig. 8. Decision boundaries for Support Vector Classifier (SVC) (a) ALL-IDB1 (b) ALL-IDB2

The data points of the two classes are visualised in Fig. 8b where a hyper plane or decision boundary is drawn in order to separate the two classes. This

hyper plane acts in support of the accuracy obtained. The hyper plane can be adjusted by changing the value of the C parameter in order to obtain a better accuracy.

7 Conclusion

In this work, an algorithm for early-stage detection of leukemia using image processing is developed. Machine learning classification techniques were used to classify between cancerous and non-cancerous cells. The proposed algorithm offers an accuracy of 95% and an approximate precision of 93% after performing image processing operations on a set of 368 blood images.

The availability of proper data set of images is necessary for better results. Getting such a uniform data set is sometimes challenging. Choosing a right method in each of the steps is essential to get accurate results. Though a lot of research is going on in this field, doctors and other biologists are reluctant on accepting and supporting this method due a chance of wrong result. This research can be extended further in order to increase the accuracy such that it can be considered as a diagnostic method for early detection of cancer.

References

1. WHO report on cancer: setting priorities, investing wisely and providing care for all. World Health Organization, Geneva (2020). Licence: CC BY-NC-SA 3.0 IGO
2. Canadian Cancer Society: Leukemia, understanding your diagnosis
3. Narayanan, U., Unnikrishnan, A., Paul, V., Joseph, S.: A survey on various supervised classification algorithms. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, pp. 2118–2124 (2017). https://doi.org/10.1109/ICECDS.2017.8389824
4. Talingdan, J.A.: Performance comparison of different classification algorithms for household poverty classification. In: 2019 4th International Conference on Information Systems Engineering (ICISE), Shanghai, China, pp. 11–15 (2019). https://doi.org/10.1109/ICISE.2019.00010
5. Agaian, S., Madhukar, M., Chronopoulos, A.T.: Automated screening system for acute myelogenous leukemia detection in blood microscopic images. IEEE Syst. J. **8**(3), 995–1004 (2014)
6. Dharani, T., Hariprasath, S.: Diagnosis of leukemia and its types using digital image processing techniques. In: 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 275–279 (2018)
7. Tran, V., Ismail, W., Hassan, R., Yoshitaka, A.: An automated method for the nuclei and cytoplasm of Acute Myeloid Leukemia detection in blood smear images. In: 2016 World Automation Congress (WAC), Rio Grande, pp. 1–6 (2016)
8. Mishra, S., Sharma, L., Majhi, B., Sa, P.K.: Microscopic image classification using DCT for the detection of acute lymphoblastic leukemia (ALL). In: Raman, B., Kumar, S., Roy, P.P., Sen, D. (eds.) Proceedings of International Conference on Computer Vision and Image Processing. AISC, vol. 459, pp. 171–180. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-2104-6_16

9. Kumar, S., Mishra, S., Asthana, P., Pragya: Automated detection of acute leukemia using K-mean clustering algorithm. In: Bhatia, S., Mishra, K., Tiwari, S., Singh, V. (eds.) Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing, vol. 554, pp. 655–670. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-3773-3_64
10. Sigit, R., Bachtiar, M.M., Fikri, M.I.: Identification of leukemia diseases based on microscopic human blood cells using image processing. In: 2018 International Conference on Applied Engineering (ICAE), Batam, pp. 1–5 (2018)
11. Choudhary, R.R., Sharma, S., Meena, G.: Detection of leukemia in human blood samples through image processing. In: Bhattacharyya, P., Sastry, H., Marriboyina, V., Sharma, R. (eds.) NGCT 2017. Communications in Computer and Information Science, vol. 828, pp. 824–834. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8660-1_61
12. Shafique, S., Tehsin, S., Anas, S., Masud, F.: Computer-assisted acute lymphoblastic leukemia detection and diagnosis. In: 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE), Islamabad, Pakistan, pp. 184–189 (2019)
13. Rege, M.V., Abdulkareem, M.B., Gaikwad, S., Gawli, B.W.: Automatic leukemia identification system using otsu image segmentation and mser approach for microscopic smear image database. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, pp. 267–272 (2018)
14. Umamaheswari, D., Geetha, S.: Segmentation and classification of acute lymphoblastic leukemia cells tooled with digital image processing and ML techniques. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1336–1341 (2018)
15. Bhagya, T., Anand, K., Kanchana, D.S., Remya, A.A.S.: Analysis of image segmentation algorithms for the effective detection of leukemic cells. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 1232–1236 (2019). https://doi.org/10.1109/ICOEI.2019.8862696
16. Kavitha, K.R., Gopinath, A., Gopi, M.: Applying improved SVM classifier for leukemia cancer classification using FCBF. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, pp. 61–66 (2017)
17. Ananya, B., Prabisha, A., Kanjana, V.: 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), Pune, pp. 1–5 (2018)
18. Donida Labati, R., Piuri, V., Scotti, F.: All-IDB website. University of Milan, Departement of Information Technologies. http://www.dti.unimi.it/fscotti/all