Alexandria Engineering Journal (2017) xxx, xxx–xxx

HOSTED BY

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

ORIGINAL ARTICLE

A decision support system for Acute Leukaemia classification based on digital microscopic images

Ahmed S. Negm [a,*], Osama A. Hassan [a], Ahmed H. Kandil [b]

[a] Department of Systems and Biomedical Engineering, High Institutes of Engineering, Al Shorouk Academy, Al Shorouk City, Cairo, Egypt
[b] Department of Systems and Biomedical Engineering, Faculty of Engineering, Cairo University, Giza, Egypt

Received 21 February 2016; revised 12 March 2017; accepted 23 August 2017

KEYWORDS

Decision support system;
Leukaemia;
Image processing;
Classification

Abstract In the era of digital microscopic imaging, Image Processing, data analysis, classification, decision support systems have emerged as one of the most important tools for diagnostic research. Physicians can observe cellular internal structures abnormalities by visualizing and analyzing images. Leukemia is a malignant disease characterized by the uncontrolled accumulation of abnormal white blood cells. The recognition of acute leukemia blast cells in colored microscopic images is a challenging task. The first important step in the automatic recognition of this disease, image segmentation, is considered to be the most critical step. In this study, we present a decision support system that includes the panel selection, segmentation using K-means clustering to identify the leukemia cells and features extraction, and image refinement. After the decision support system successfully identifies the cells and its internal structure, the cells are classified according to their morphological features of this analysis the decision support system was tested using a public dataset designed to test segmentation techniques for identifying specific cells, and the results of this analysis were compared with those of other techniques, which were suggested by other researchers, applied to the same data. The algorithm was then applied to another dataset, extracted under the supervision by an expert pathologist, from a local hospital; the total dataset consisted of 757 images gathered from two datasets. The images of the datasets are labeled with three different labels, which represents three types of leukemia cells: blast, myelocyte, and segmented cells. The process of labeling of these images was revised by the expert pathologist. The algorithm testing using this dataset demonstrated an overall accuracy of 99.517%, the sensitivity of 99.348%, and specificity of 99.529%. Therefore, this algorithm yielded promising results and warrants further research.

1. Introduction

All blood cells arise in the bone marrow, which occupies the central cavity of the bone, via hematopoiesis. Specifically, stem cells differentiate into specific blood cells [1], such as white blood cells, which include neutrophils, monocytes, eosinophils,

* Corresponding author.
E-mail address: ahsnegm@gmail.com (A.S. Negm).
☆ Acute Leukemia Images classified according to digital image processing and pattern recognition.
Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2017.08.025

basophils, and lymphocytes [2]. Leukemia refers to cancer in blood cells, and the bone marrow generates leukemia cells, which are abnormal white blood cells. Acute leukemia causes the fast deterioration of the patient, whereas chronic leukemia is characterized by slow progression and may be lymphocytic or myelogenous. In addition to these subtypes, leukemia can be classified into the following types: Acute lymphoblastic (ALL), Acute myelogenous (AML), Chronic lymphocytic (CLL) and Chronic myelogenous (CML) [2].

Currently, leukemia has been classified using two systems: the French-American-British (FAB) classification and the World Health Organization (WHO) proposal [2]. Blast cells found in the peripheral blood smear and characterized Acute Myeloid Leukemia (AMLs), which includes seven types (M-1–M-7). Hematologists microscopically examine the blood under a light microscope. This process is very tedious, time-consuming and not suitable for analyzing a large number of cells. Nevertheless, some mathematical approaches and technologies have been developed to discriminate blood cells, and the picture-preparing stage is essential for artifact extraction and identifying leukemia [3].

Aimi Salihah et al. [3] they state that, in order to investigate an image, the first important step is its division to generate regions with significant value and less demanding ones to break down.

According to Daniela et al. [4] segmentation enables artifacts to be separated without the incorporation of non-essential material by defining the boundaries of the blood cells. The similarity is the key step to arranging image pixels into locales that compare semantically significant substances during segmentation.

Two forms of segmentation have been described, pixel-based image segmentation and region-based segmentation. Zuva et al. [5] have stated that the shape segmentation techniques are classified into threshold-based, edge-based and region-based techniques.

Thresholding and clustering have been considered the simplest segmentation techniques [6,7]. Thresholding is suitable only for the items that do not touch one another and whose depth markedly differs from that of the background [6]. Clustering has been widely used to segment grey-level images [7,8]. Kim et al. [6] have considered threshold, edge detection, pixel clustering, and region growing are examples of segmentation techniques. They have been used together to extract the nucleus and cytoplasm of leukocytes [9,10]. Prasad et al. [8] have used the auto-segmentation of blood cells for counting. The threshold, chessboard distance measure, and watershed have been used for the segmentation of blood cells. Chen et al. [10] have noted that the watershed segmentation algorithm yields a good result with the distance transform.

Kekre et al. [11,12] have stated that a good codebook is a key to Vector Quantization (VQ). The Linde-Buzo-Gray (LBG) algorithm, also called the Generalized LIoyd Algorithm (GLA), and K-means cluster is commonly used methods to generate a codebook. In limited cases, techniques, such as isodata, fuzzy c-means, and k-means, have been applied to color images [12–14]. Minimizing the sum of squared distances between all points and the cluster centre are the K-means.

The problem of overlapped cell has been overcome by splitting cells via joining concave points by separating lines. Dorini [7], and Hengena et al. [15] have used eroding and region growing for regions retaining the shape. For automated cell splitting, the watershed technique has been used to differentiate the clustered cells to improve tallying [15]. Yan et al. [16] have split the overlapped objects by combining the distance transform and watershed algorithm. Labati et al. [13] have used morphological operators to process data based on the characteristics of the objects and their shapes in the input image, which is encoded in the structuring element. The morphological operations and median filters have been suggested to be used only for noise removal.

Recent studies described the morphological feature analysis of lymphoma and leukemia cells, such as that by Schäfer et al. [17] for Hodgkin's lymphoma. Standard pre-processing methods, such as Gaussian filtering, the application of a threshold for background elimination and region labeling for the identification of relevant tissue patches, are applied. After pre-processing, the pixels considered to represent tissue, are classified using a supervised approach. For each pixel class, the relative fraction of pixels belonging to the respective class was determined.

Mohamed et al. [18] have presented a white blood cell nucleus segmentation algorithm. Their proposed algorithm is based on the Gram-Schmidt orthogonalization technique. The orthogonalization technique processes the RGB colors. It enhances the arbitrarily selected color and diminishes the two other colors. They demonstrate the highest contrast for the nucleus. Morphological operations were used to enhance the segmentation. Mohammed et al. [19] presented a method to segment normal and CLL lymphocytes into two parts, the nucleus, and cytoplasm, using a watershed algorithm and optimal thresholding.

Madhukar et al. [20] have stated that some halfway/fully mechanized frameworks for leukemia are still at the model stage. Presently, five primary peculiarities are utilized via mechanized frameworks for ahead-of-schedule ALL recognition: cell size, shading, shape, thickness, and granularity. Both the sophisticated nature of the blood pictures and the variable slide planning systems are considered to make an appropriate clinical decision [21]. Additionally, most current frameworks focus on artifacts in the sub-pictures rather than the complete blood smear. A few frameworks have been proposed as systems to refine the division and to avoid the incorrect segmentations of white cells.

Madhukar et al. [9] have followed four fundamental handling steps [20]. Pre-processing of the image, division of the entire image, determine distinctive arrangements of peculiarities for a database of images, use the classifier framework. The Support Vector Machine (SVM) classifier result in an accuracy of 93.5% on 98 separate arrangements of images.

Sadeghian et al. [22] have stated that the utilization of picture-preparing systems have grown quickly in recent years and can provide data about the degree of core versus cytoplasm to distinguish and characterize distinctive White Blood Cells (WBCs), which actualizes automatic thresholding technique proposed by Otsu (8). A normal exactness of 92% for the core division and 70% for cytoplasm division was reported by Belsare [23]. Histopathology refers to the examination of intrusive or less obtrusive biopsy tests by a pathologist under a magnifying instrument for placing, dissecting and grouping the majority of ailments, such as tumors.

Researchers have suggested that models based on decision trees can be used to make decisions about the treatment of glaucoma patients. Kumar et al. [24] developed decision-support

Please cite this article in press as: A.S. Negm et al., A decision support system for Acute Leukaemia classification based on digital microscopic images, Alexandria Eng. J. (2017), https://doi.org/10.1016/j.aej.2017.08.025

systems for diabetes, hepatitis and heart diseases to help physicians. Specifically, they proposed an algorithm to classify these diseases and compared the effectiveness, correction, C4.5 better than an ID3 algorithm with overall accuracy 71.4% [24]. Soni et al. [25] developed a predictive data-mining system for heart disease prediction based on three distinctive regulated machine learning calculations, Nave Bayes, K-Nearest Neighbors (K-NN), and Decision tree [25]. Artificial Neural Networks (ANNs) are currently being utilized to handle several problems related to display due to their unique properties, like their non-parametric nature, non-linear nature, and ability to conduct input-output mapping. ANNs have also been used to distinguish a specific pathology, such as tumour analysis, the programmed distinction of readiness and languor from electroencephalography, forecasts of coronary course stenosis, the examination of Doppler movement flags, the characterization and prediction of the movement of thyroid-related ophthalmopathy, diabetic retinopathy grouping, saccade discovery in EOG recordings and PERG characterization [26].

Huang et al. [27] described that computer vision methods have generally been utilized during diagnosis over the past decade. A calculation based k-means classifier is used to characterize leukocytes into five categories. The exploratory results demonstrate that the distinction rate obtained using the LNSC reaches 91%. According to Pandey and Mishra [28], ANN is an information-prevailing methodology that is generally utilized as a part of the medical field, differentiated by using both the learning law and topology. ANN has been successfully used in different medical applications in many areas of medicine, as illustrated in a study conducted in Lisboa [29] This approach has some disadvantages, such as the structure of NN, which is not straightforward. These researchers surmised a subjective discovery model of the mapping standard, and from the earlier master learning cannot be considered to better ordaining the network parameters, keeping in mind that the end goal is to enhance meeting and decrease the learning time.

Some of the applications included in our study are the response to HP eradication recurrence, lymph node metastasis, response to interferon in chronic hepatitis C, and diagnosis of diabetes occurrence. Galindo [30] proposed trials sufficient to investigating the core, cell, and cytoplasm. We followed this approach in this work by utilizing only the peculiarities of entire cells. In every analysis, built diverse grouping errand to recognize among sorts of intense Leukemia (ALL and AML, subtypes of ALL (L1 and L2), and subtypes of AML (M2, M3, and M5).in the instance of AML subtypes we examined the conduct of a subtype concerning the others performing the paired and multiclass orders [31]. For every arrangement issues, utilizing distinctive sorts of gimmicks, for example, geometric, measurable, surface, and size proportion, the order was done utilizing serious based classifiers, choice trees, lament particle works and in addition meta-classifiers accessible in (weka 2009) [31].

We herein present bone marrow images with heterogeneous staining and pixels features, such as color and texture, which our segmentation algorithm based on. The remainder of the manuscript is organized as follows: In Section 2, the proposed algorithm is explained in the materials and methods. In Section 3, the results and discussion obtained using the algorithm are presented. In Section 4, the conclusions is presented.

2. Materials and methods

The proposed algorithm is shown in Fig. 1

2.1. Material

We have applied our proposed system to the datasets as follows:

- Our system was applied and tested on 115 digital light microscopy images 632 * 480 pixels in size. This dataset was also used by Kekre [12].
- Data were collected using a light microscope supported by CCD camera. Six hundred and twenty-four images that were 632 * 480 in size were classified into three categories by an expert.
- The data were processed using a laptop with 2 microprocessors (Centrino Mobile Technology), 1 GB RAM and 256 MB Vega with the Microsoft Window XP operating system and Programming language MATLAB 7.

2.2. Methods

We aimed to generate a successful algorithm that identifies cancerous blood cells.

To this end, we review a range of important points before describing the algorithm:

- To ensure that blast cells could be identified in the public dataset were compared the result with those of other researchers.
- We expanded the dataset by adding another dataset containing three other types of leukemia cells: blast, myelocyte, and segmented cells, to expand the scope of recognition by the algorithm and test its sensitivity and specificity.
- The large local dataset was used for the more sophisticated stages to export the cell features, and this information was used to feed the classification step to automatically differentiate leukemia cells.
- We excluded the public dataset from being used in the advanced process of the algorithm because it was not established for this purpose and it limited the ability of the segmentation step to identify only cells, as authorized by the owner.

2.3. Preprocessing

The images were represented by three color components, RGB. The three color component images, red, green and blue, were visually tested and are shown in Fig. 2. The histogram of green color distribution presented below in Fig. 2 indicates that the green component contains the most contrast information. We selected this plane for the subsequent segmentation step.

2.4. Processing

The proposed method is based on the k-means algorithm, which begins by partitioning a vast arrangement of vectors

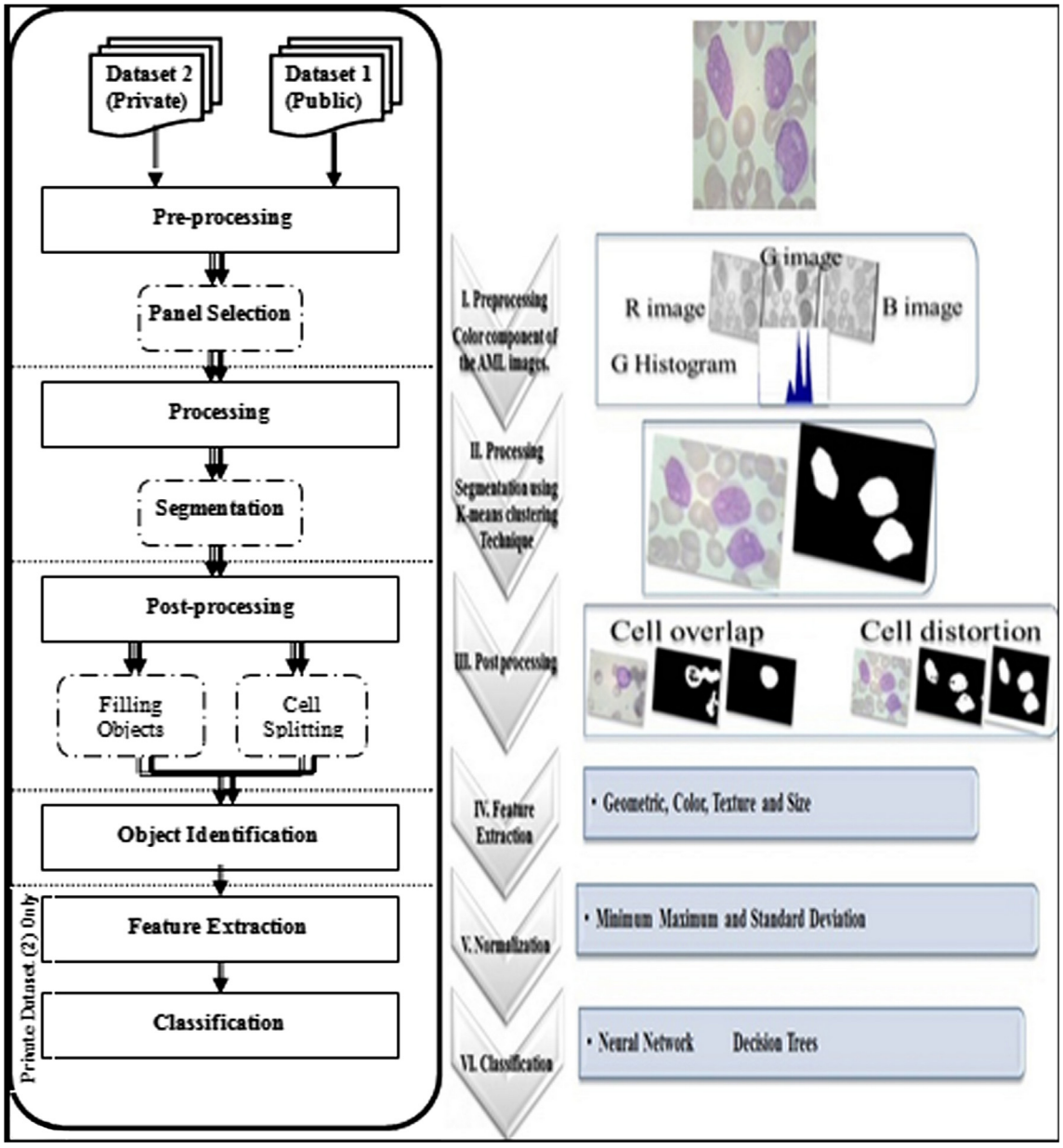


Fig. 1 Consecutive sequential steps that illustrate and guarantee the achievement of the objective of the algorithm.

into groups having the same number of points. The centroid point represents the group. By partitioning an image into K-clusters, classifying and grouping items into k groups (k is the number of pre-selected groups), minimizing the sum of squared distances between items and the corresponding centroid used in grouping the segmentation performed [32]. The procedure consists of the following steps:

1. K initial cluster centers Z_1 (1)…Z_K are selected.
2. Distribution of samples {x} among the K clusters using the relation by the kth iterative step.

$$X \epsilon C_j(K) \text{if} \|X - Z_j(K)\| < \|X - Z_i(K)\| \quad (1)$$

for all i = 1, . . . , K; i j; where Cj (k) denotes the set of samples whose cluster centre is $z_j(k)$.

3. New cluster centers, Z_j (k + 1), j = 1, 2, . . . , K are computed. Minimizing the new cluster according to the sum of the squared distances from all points in C_j (k). Minimization is measured by the mean of C_j (k).

$$Z_j(K+1) = 1/N_j[X \epsilon C_j(K) \sum X, j = 1 \cdots K] \quad (2)$$

Z_j (K + 1) is new cluster centre where N_j is the number of samples in C_j(k).
4. If Z_j (k + 1) = Z_j (k) for j = 1, 2, . . . , K, the algorithm terminates. Otherwise, it goes to Step 2.

The value of K initial cluster centers used in clustering is specified as input to the algorithm. We applied a K-means clustering algorithm with k = 3 followed by K = 2 for segmenting the blast cells. In each step, the returned clustered pixels are the maximum green value.

Please cite this article in press as: A.S. Negm et al., A decision support system for Acute Leukaemia classification based on digital microscopic images, Alexandria Eng. J. (2017), https://doi.org/10.1016/j.aej.2017.08.025

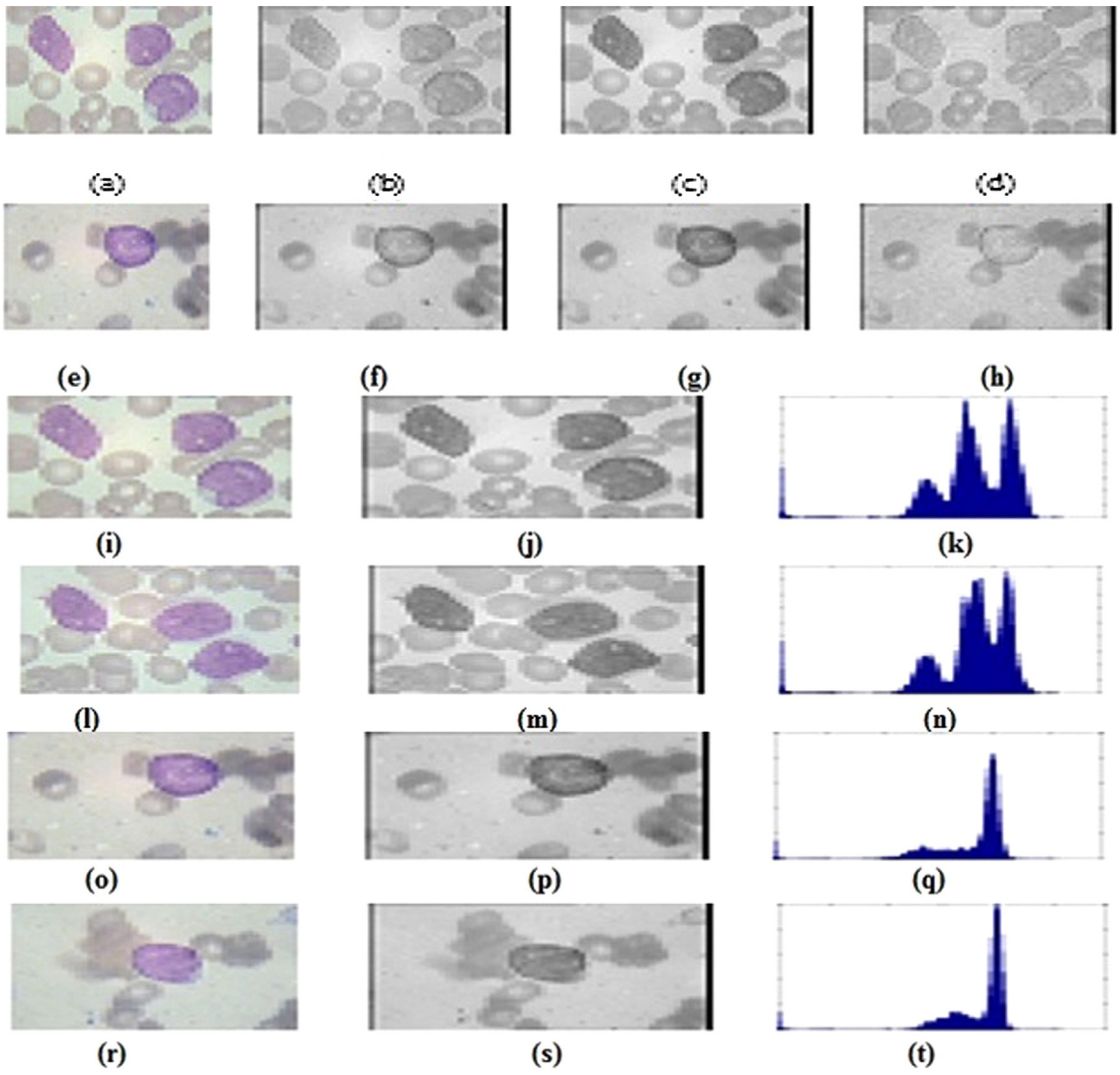


Fig. 2 Sample of a color component of the Acute Leukemia images (a and e) Original Image. (b and f) Red color component of the image. (c and g) Green color component of the image. (d and h) Blue color component of the image. Color image of Acute leukemia with its green component image and the green color distribution histogram; (i, l, o, and r) Acute Leukemia images, (j, m, p, and s) Green color component image and (k, n, q and t) Histogram of green color distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The following Matlab sentences are a function that we used to apply the K-means clustering step.

```
[idx  cent] = kmeans(double(dta),3,'start','uniform','Emp
tyAction','drop');%'start',
[idx2  cent2] = kmeans(double(dta(object,:)),2,'start','uni
form',EmptyAction','drop');
```

We applied a K-means clustering algorithm to separate the desired cells successfully been through two steps as follows: 1-Dividing green intensity components of the image into three classes, each class represents a component of the image components (background, other non-target cells, the cells to be extracted) by applying the K-means algorithms with K = 3 to determine the centroids of segmented green channel the maximum. Retrieve of the highest green color intensity which expresses the desired cells. 2-Re-divide the former green component retrieved from the previous step and representing the cells and internal components into two parts, which helps to achieve the separation of the cells and the nucleus and to distinguish between them. Fig. 3 demonstrate these step.

2.5. Post-processing

Unwanted regions are present in the segmented image. Post-processing is required to allocate the blasts only. Image problems, such as cell overlapping and cell distortion, are solved in the enhancement step. We have proposed a cell separation algorithm that maintains the original shape of the blood cell and uses information on its shape to split the overlapped regions in Fig. 4. Pixels are associated into groups based on the variance using minimum variance quantization [13]. Identifying the overlapped objects as a single object leads to errors in measurements and statistics. We addressed this problem by

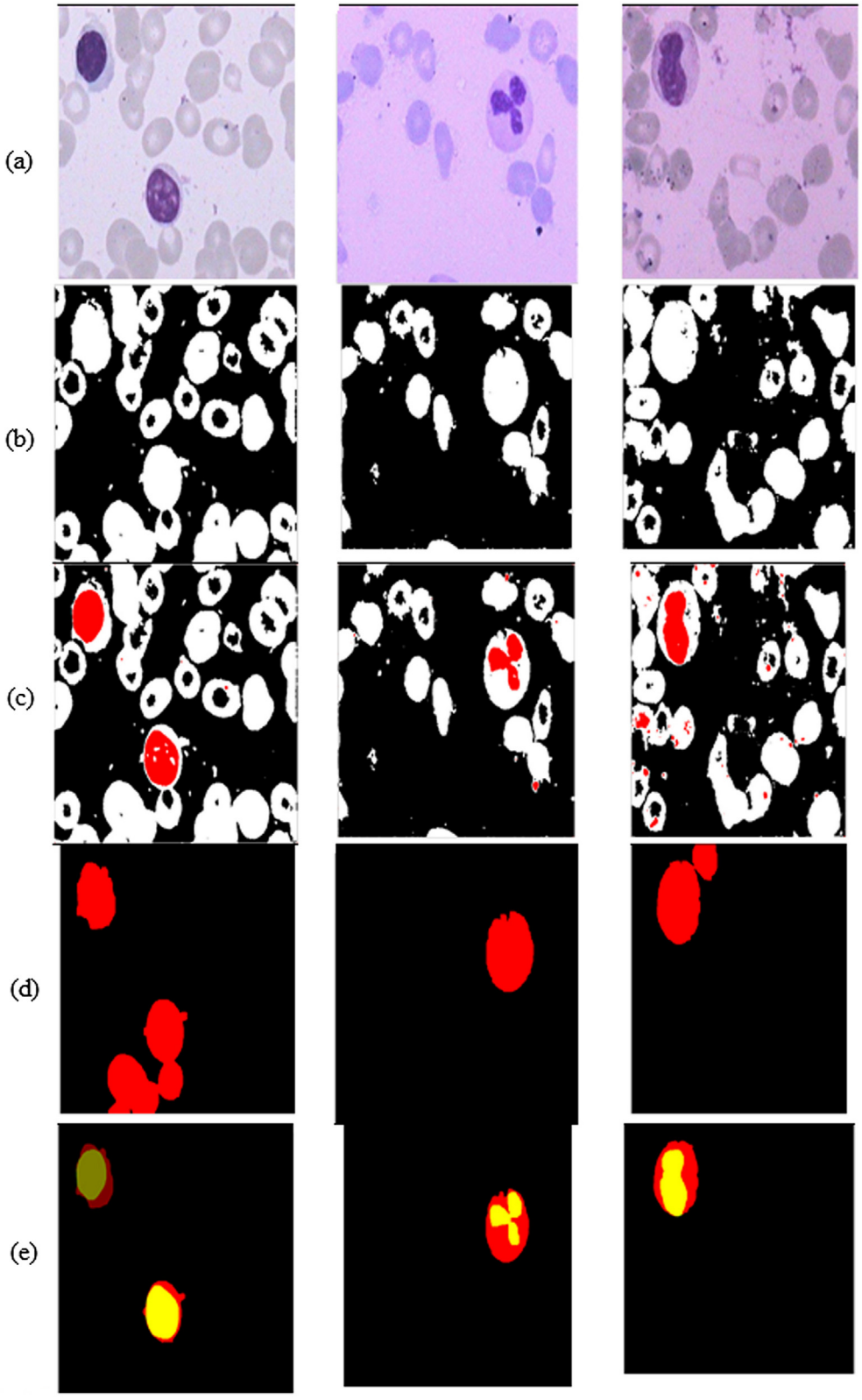


Fig. 3 Shows the application of the former algorithm steps dependence on the new database collected by us include: (A) selection of images of the different constituent cells of the database. (B) The application of segmentation steps for the cells and identify them. (C) Identify the components of cells and their internal contents "nuclei". (D) Separating the desired cells from any impurities specially overlapped cells and distorted one. (E) The final status of the image after the exclusion of all posters and defects.

Please cite this article in press as: A.S. Negm et al., A decision support system for Acute Leukaemia classification based on digital microscopic images, Alexandria Eng. J. (2017), https://doi.org/10.1016/j.aej.2017.08.025

watershed distance segmentation. This method automatically splits the overlapped objects [10].

The following Matlab sentence is function of watershed which we use L = watershed(D);

- Erosion [34]:

$$A \ominus B = \{z \in E | B_{\approx} \subseteq A\} \tag{3}$$

- Dilation [34]:

$$A \oplus B = \bigcup_{b \in B} A_b \tag{4}$$

- Closing [34]:

$$A \bullet B = (A \oplus B) \ominus B \tag{5}$$

- Opening [34]:

$$A \circ B = (A \ominus B) \oplus B \tag{6}$$

We use an opening operator for fill the missing pixels in the cells according to their similarity to the background color, as shown in Fig.4. The following Matlab sentences are functions we use.

```
bin3 = imfill(bin3,'holes');
bin3 = imopen(bin3,strel('disk',7,8));
bin3 = bwareaopen(bin3,800);
```

2.6. Data normalization

We used the following normalization techniques to compensate the differences in the scales of the chosen features.

- *Min*—max *normalization* [7]:

$$\mathcal{V}'(i) = (v(i) - \min(v(i)))/(\max(v(i)) - \min(v(i))) \tag{7}$$

- *Standard deviation normalization* [7]:

$$\mathcal{V}'(i) = \frac{v(i) - mean(v)}{sd(v)} \tag{8}$$

2.7. Features extraction

The Features extraction facilitates both the classification and recognition of leukemia cells. In this step, the extracted features are based on the geometry, statistics, textures, and size ratio from regions selected in the segmentation process (nucleus, cytoplasm, and whole cell). An analysis of these features is then performed to differentiate the types and subtypes of acute leukemia [35].

All aforementioned geometric features were extracted from each nucleus and cell in Table 1. Features represent nucleus and cell because of the cell contain cytoplasm feature also. And we calculate the area for cytoplasm only. Due to the morphological analysis perform by the expert verify that cytoplasm features are not relevant to the classification of cells. In the case of the nucleus, cytoplasm, and cells, we extracted statistical and textural features from the channels of the RBG image and the grey-scale image [35].

These features were analyzed to classify acute leukemia cells using different training and testing sets, attribute selection, and classification algorithms available in Waikato environment for knowledge analysis version 3.7.9 (Weka). Weka produced by the University of Waikato, Hamilton, New Zealand; is open source software issued under the GNC General Public License, it is a collection of machine learning algorithms for data mining tasks. The algorithm can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [36].

2.8. Decision tree

Decision tree learning is one of most broadly utilized and basic routines for inductive derivation. It a strategy for approximating discrete-esteemed target works in which the educated capacity is spoken to by choice decision tree.

Choice decision trees order occasions by sorting them down the tree from the root to a leaf hub, which arranges occurrence. Every hub in the tree determines a test of some characteristic of the occasion, and every extension sliding from the hub is

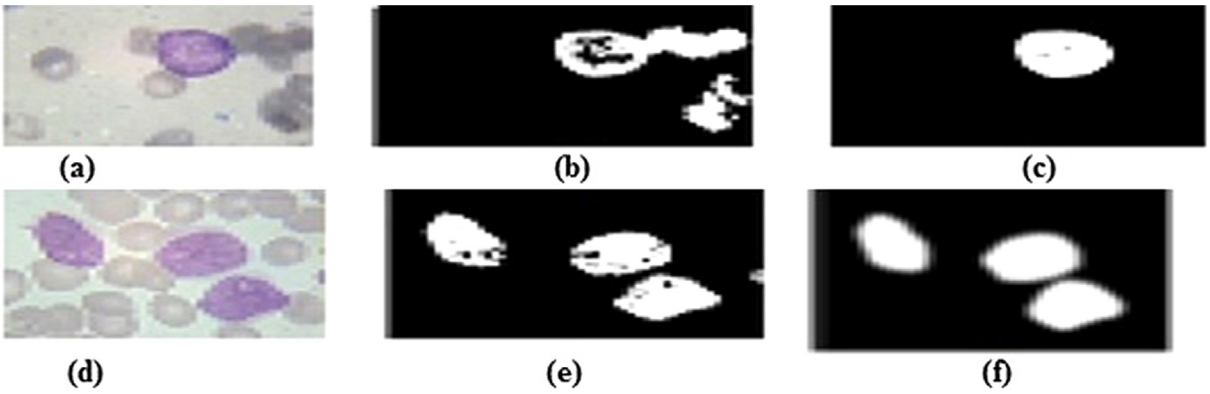
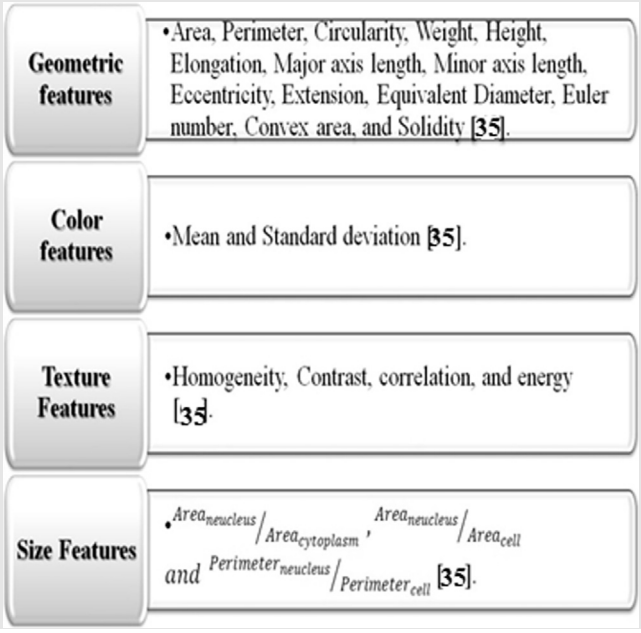


Fig. 4 Shows overcome the segmentation problems as cell overlap and cell distortion; (a and d) are original images feed the algorithm. Cells overlap splitting and extract the cell shown in (b) which present segmented image with overlapping and (c) shows the succession of cell extraction that's happened by using distance transform combined with the watershed algorithm. Segmented images with cell distortion and morphological operator filling gaps (e) Segmented image with cell distortion and (f) cell after filling gaps.

Table 1 Extracted features from Acute Myeloid Leukemia cells [35].



compared to a conceivable quality for this trait. An occasion is grouped by beginning at the root hub of the tree, testing the trait determined by this hub, and subsequently down the tree limb by a comparison to the estimation of the characteristic; this methodology is then rehashed for the sub-tree established at the new hub [37].

This approach is exemplified in C4.5. [32] The decision tree starts learning by constructing trees top-down. We characterize a measurable property called data to pick up (IG), which measures how well a given characteristic divides the preparation information based on its target characterization.

$$IG = E_{before} - E_{after} \quad (9)$$

where E is the entropy, which measures the uncertainty associated with a random variable; it describes the (im) virtue of a self-assertive gathering of data. At each node of the tree, this calculation is performed for each feature, and the feature with the largest IG is selected for the split; this process continues iteratively until the end. C4.5 is a calculation used to create a Decision tree grown by Ross Quinlan. C4.5 is an augmentation of Quinlan's prior ID3 calculation. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. 298 C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can 299 be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. It 300 became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent 301 paper published by Springer LNCS in 2008 [32]. C4.5 is inherent model Weka. Our tree were 302 shown in Fig. 5.

2.9. Artificial Neural Network (ANN)

Multilayer perceptron algorithms are often used for classification problems in medicine. All MLPs consist of 2 concealed layers with digression hyperbolic exchange capacities and a yield layer of single neurons with a logistic exchange work that give the MLP yield. A number of neurons in the shrouded layers was selected by considering the type of data utilized as a part of a request to attain to the best execution judged based on the outcomes of the 10-fold cross-acceptance technique. ANNs were prepared and tested with the 10-fold cross-approval strategy to diminish to reduce the possibility of testing the same sample, while completely using our information set. Information was arbitrarily partitioned into ten subsets. One subset was utilized to test grouping execution, whereas the remaining nine subsets were utilized for preparing purposes. In our ANN, one of the nine preparation subsets was retained for the right-on-time ceasing of the ANNs, keeping in mind that the end goal was to avoid over-fitting. We use Weka classifiers functions Multilayer Perceptron for Neural Network classification. The inputs of the network consist from 648 instances of attributes 23 features extracted from the cells shown in Table 1. The output of network attributes no 24 is a grade of each cell classified into three classes of cells Myeloblast, Myelocyte and segmented as shown in Fig. 6. Taken time to build model is 2.47 s.

3. Results and discussion

Because one of the main objectives of this work was to objectively rather than subjectively identify blast cells, we had to define specific limiting parameters before effectively applying our suggested algorithm. The algorithm was applied to a well-known public dataset of blood samples organized by Dr. Fabio Scotti to its segmentation and image classification performances [13]. This dataset was also used by Kekre [12]. The algorithm consists of pre-processing, processing and post-processing, and these main steps are agreed upon among many researchers in the field [12,17].

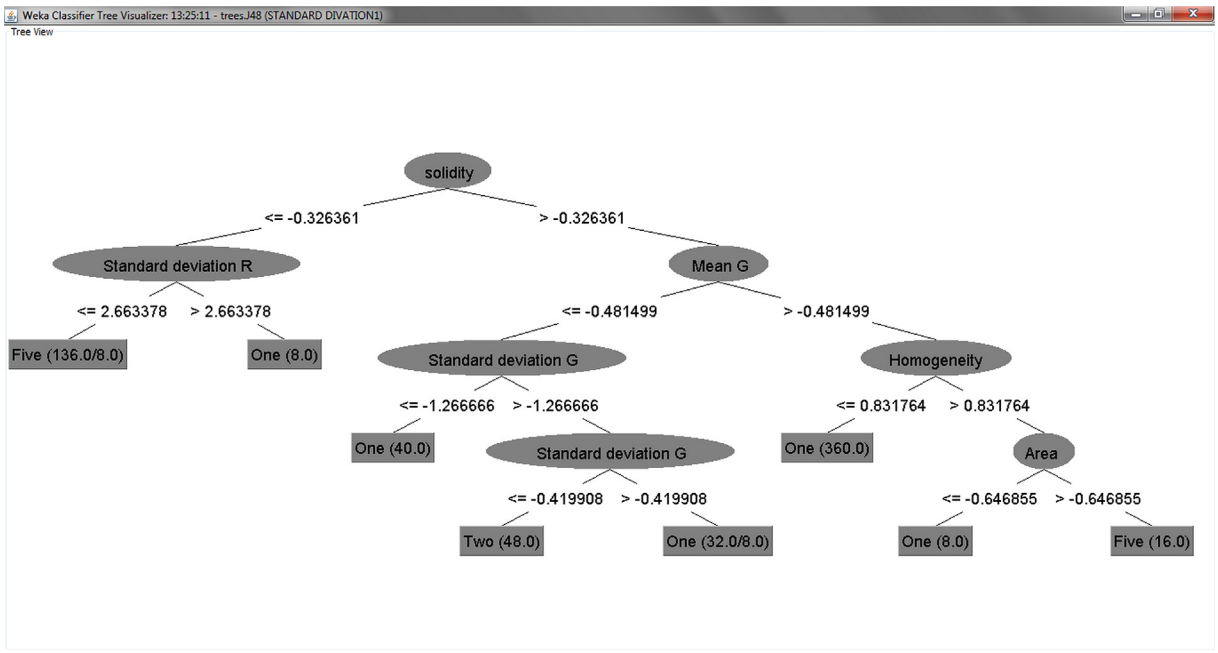


Fig. 5 Representation of the Decision Tree from the root to a leaf hub showing information gain to features to classify the three types of cells.

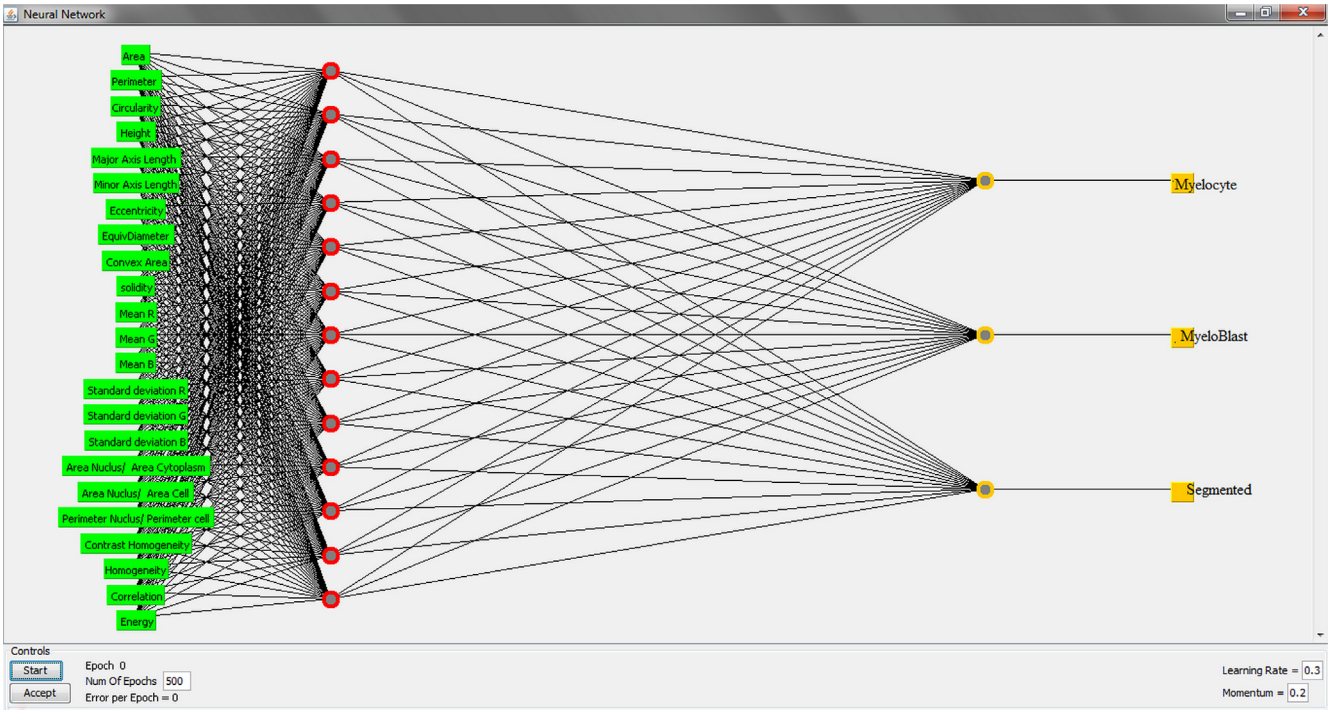


Fig. 6 Shows inputs of the neural network consist from 648 instances of attributes 23 features extracted from the cells. The output of network attribute is a grade of each cell classified into three classes of cells Myeloblast, Myelocyte, and segmented cells.

The developed algorithm is based on the panel selection, segmentation using K-means clustering, and refinement processes to detect abnormal white blood cells (blast) [13,32]. Pre-processing is a critical step that ensures the success of all subsequent algorithm steps. Whereas the image processing technique is subject to selection, e.g., Schäfer et al. used Gaussian filtering for Hodgkin's lymphoma images in [17], a threshold must be applied to eliminate background. Kekre [12] used the same dataset used herein, which was separated into three color planes, red, green and blue. They used both the green

and blue planes, whereas we relied on only the green plane because the green component contained the highest contrast information Fig. 2.

Segmentation step is the cornerstone of any algorithm. Various vector quantization techniques are used for image segmentation, as shown by Kekre et al. in [12], including LBG, KPE, and K-mean clustering, as demonstrated in this study Fig. 7.

The outcome of segmentation step is represented by specific features that need to be addressed as follows: overlapped cells and the same color density of cytoplasm and background, not fully connected contour.

Cell splitting can be used to overcome the problem of overlapping using automatic or manual techniques [8,15]. We applied automatic cell splitting using watershed transform followed by a morphological operation to overcome overlapping and cell splitting. The success of identifying blast cells in digital microscopic images after completion algorithm processing is based on overcoming all problems that have been encountered in the images.

Concerning such implements as an algorithm for successful detection of blast cells, the accuracy of the algorithm was carefully evaluated to estimate its performance. The specificity and sensitivity of the suggested strategy were assessed using the following formula:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) = 1 - \text{FPrate} \tag{10}$$

$$\text{FP rate} = \text{FP}/\text{N} \tag{11}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \tag{12}$$

The overall Accuracy calculated by

$$= (\text{TN} + \text{TP})/(\text{TN} + \text{TP} + \text{FN} + \text{FP}) \tag{13}$$

where (TP) stands for True Positive and measures the number of blast cells effectively recognized as blast cells; (TN) stands for True Negative and measures the number of non-blast cells correctly identified as non-blast cells; (FP) stands for False Positive and measures number of cells falsely identified as blast cells; (FN) stands for False Negative and measures the number of cells falsely identified as non-blast cells.

According to the suggested strategy described above, we need to ensure that the algorithm can identify leukemia cells using the public dataset. Table 1 illustrates the confusion matrix of applying the algorithm to the public dataset. Of the 2319 cells, 265 were identified true positive segmented blast cells. According to the data represented in this table, the sensitivity and specificity of the algorithm were 97.4% and 98.1%, respectively, with an accuracy of 98.06%.

This public data set was applied to our algorithm, which depends on the K-means segmentation technique. Fig. 8 compares the successful identification of the cells in Acute Myeloid Leukemia (AML) images or Acute Lymphoid leukemia (ALL) images against two other techniques described by Kekre, which use LBG and KPE vector quantization [12]. This comparison shows that our algorithm is superior to the other two techniques, as demonstrated by the sensitivity and specificity of identifying blast cells. Accuracy Acute Leukemia cell

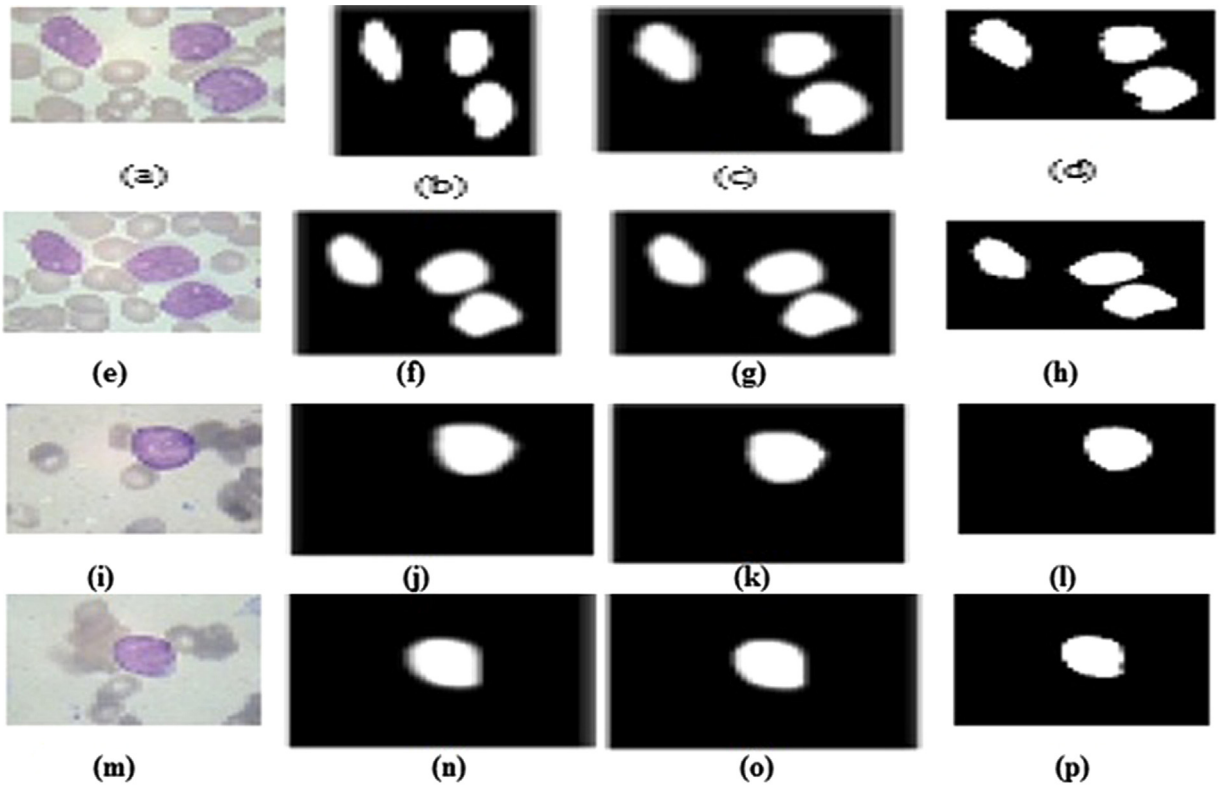


Fig. 7 Different vector quantization algorithms applied on Acute leukemia images from public dataset (a and e) Acute Myeloid Leukemia Images. (i and m) Acute Lymphoblastic Leukemia images. Images (b, f, j, and n) are after applying Linde-Buzo-gray LBG [12]. (c, g, k and o) Images after applying Kekre's Proportionate Error algorithm KPE [12]. (d, h, I and p) Images after applying K-means algorithm.

detection after applying the algorithm in each dataset were demonstrated in Fig. 9.

As mentioned above, the scope of the algorithm was expanded in the identification verification step by applying it to a new local dataset. This set contains three different types of cells: blast, myelocyte and segmented cells, which had previously been labeled by an expert and extracted from 624 Acute Myeloid Leukemia images Fig. 8. Table 2 shows that the algorithm successfully identified 802 cells with a sensitivity and specificity 100% and 99.747%, respectively, and an accuracy of 99.76%.

Thus, the algorithm was evaluated based on its sensitivity and specificity for two different datasets, yielding impressive results. The recognition of acute leukemia blast cells in colored microscopic images segmentation by three steps pre-processing "panel selection", processing segmentation using K-means clustering to identify the leukemia cells and features extraction "image refinement". Merging data obtained from both sets showed that the algorithm identified three types of leukemia cells in the 757 images. These cells were identified with an accuracy of 99.709%, sensitivity of 99.348% and specificity of 99.707%.

After the algorithm was shown to be able to identify various types of acute leukemia cells, it was refined to automatically classify these cells. To this end, a local dataset was used. The cell features were exported to an excel sheet used to feed the classifiers. Features that represent each cell shown in Table 1 were divided into geometric, color, texture and size features.

According to the wide range of numerical representation number, the features prepared by the normalization step before applying to the classifiers. We use two normalization techniques, minimum-maximum and standard deviation, and two classifiers, neural network, and decision tree. Thus, four combinations were used to test the influence of normalization on the classifying step. Table 3 demonstrates that the quantization technique does not affect the result either classifier. Thus, the classifier output only depends on the classifier.

The outputs derived from these descriptions were examined and subjected to a statistical analysis. Both methods used in the unification highlighted that the results did not affect the classification step.

The final phase of the algorithm is the classification of data using previously clarified in terms of the number and configuration base. The images were then exposed to the proposed algorithm. A pathology expert (Prof. Dr. Osama Hasssan, one of the authors of this work) had already accessed the data and classified the three types of cells.

The insert images and outputs of the cells to the two methods of classification, Neural Network, and Decision Tree, were used to determine the degree of accuracy and transparency of each method of classification, as shown in Table 3.

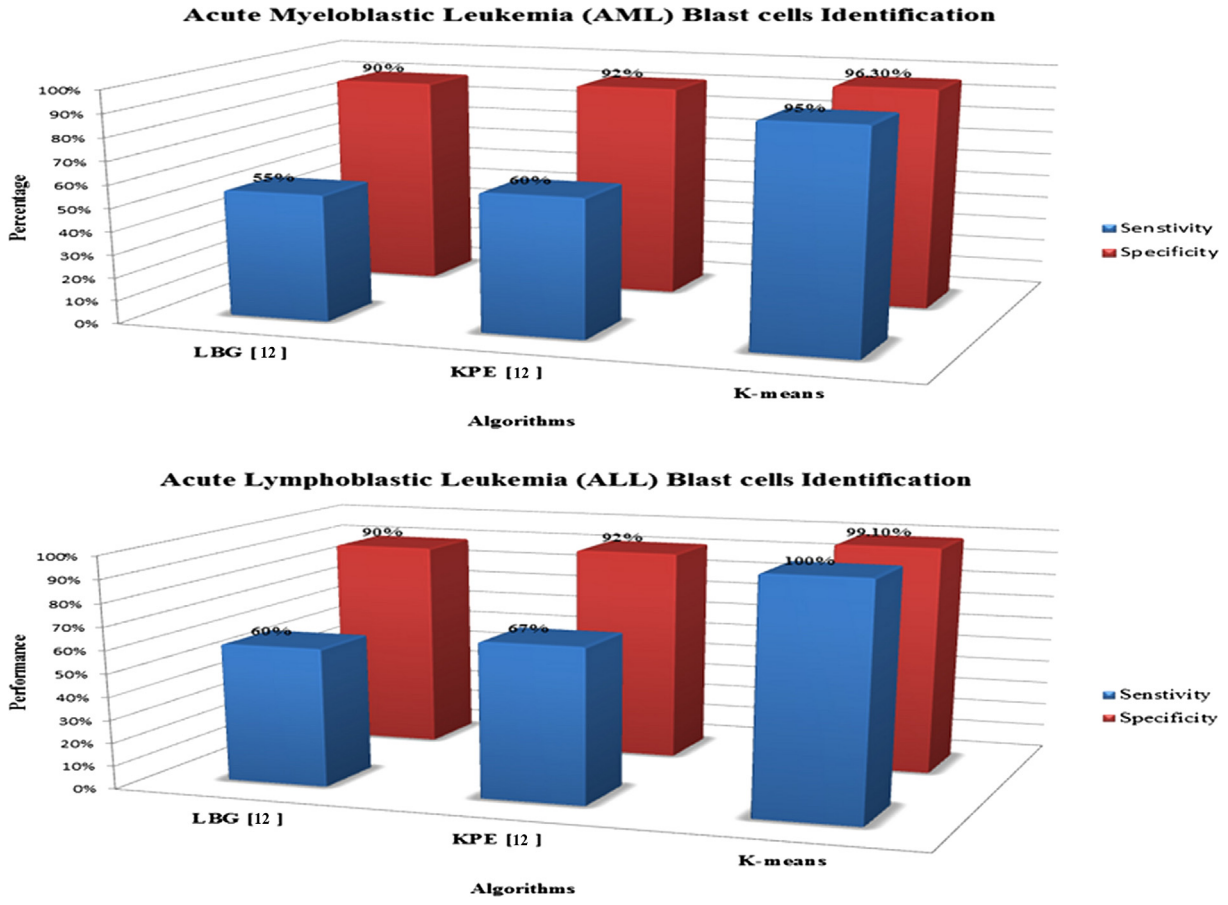


Fig. 8 Comparison between three different algorithms implemented to identify blast cells in Acute Myeloid leukemia (AML) and Acute Lymphoid Leukemia (ALL) images of public dataset [12].

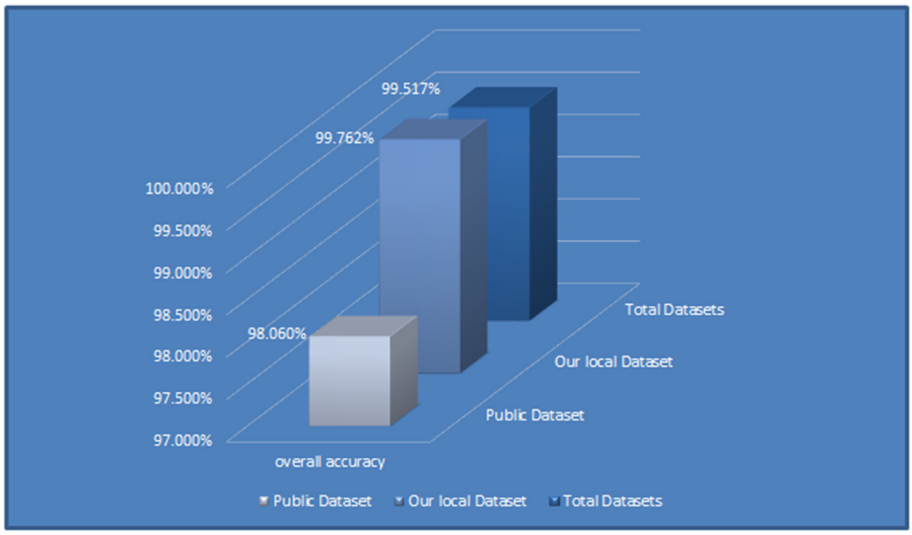


Fig. 9 Accuracy of applying the algorithm to each data set.

Table 2 Confusion Matrices of blast cells identifications for both public and local datasets.

Public dataset			Our local dataset		
No. of images 115	Predicated acute leukemia cells		No. of images 642	Predicated acute leukemia cells	
Actual stages	True	False	Actual stages	True	False
Positive	265	38	Positive	802	33
Negative	7	2009	Negative	0	13013
Overall accuracy	98.06%		Overall accuracy	99.7617%	

Table 3 Comparison between two Confusion Matrices ensure algorithm ability to differentiate three group of cells depend on its extracted features by two classifiers neural network and decision trees.

True stages	Predicated cells					
	(Neural network)			(Decision tree)		
	Myelo cell	Blast cell	Segmented cell	Myelo cell	Blast cell	Segmented cell
Myelo cell	91%	9%	0%	88%	13%	0%
Blast cell	1%	98%	1%	1%	97%	2%
Segmented cell	0%	3%	97%	0%	2%	98%
Overall accuracy	96.76%			96.60%		

Ten cross-validation test techniques were used to build each classifier model. The neural network is more sensitive than decision tree in classifying the three leukemia cells in the dataset. Neural network correctly classified 627 of 648 cells (96.75%) and failed to classify 21 objects (3.24%). Although the decision tree cannot keep pace with the neural network in classifying myelocyte cells and blast cells, with identification rates of 88% and 97%, respectively, it is more sensitive in differentiating segmented cells, with a sensitivity of 98%. Furthermore, the decision tree model was faster than the neural network; it requires 0.03 s, whereas the neural networks require 2.47 s.

The overall accuracy did not significantly differ between the classifiers. Based on the acceptable run time of the neural network model, we suggest using this technique due to its ability to differentiate each group of cells.

4. Conclusions

Microscopic digital images have been analyzed to identify blast cells in leukemia. The algorithm can identify blast cells under specified criteria of image processing and enhancement.

Image segmentation was applied to automatically record the frequency of specific repetitive objects (cells). Image enhancement was applied to improve the image quality.

The proposed algorithm, which uses K-means clustering, was compared with LBG and KPE for blast detection in acute leukemia images. The K-means algorithm was superior to the LBG and KPE algorithms.

This system was applied to a second database, which consists of 642 images. This application proved that the algorithm successfully discriminated cells with a sensitivity of 100% and

Please cite this article in press as: A.S. Negm et al., A decision support system for Acute Leukaemia classification based on digital microscopic images, Alexandria Eng. J. (2017), https://doi.org/10.1016/j.aej.2017.08.025

an accuracy of 99.74%. Thus, leukemia cells can be characterized based on the following:

- Geometry characterization.
- Characterization of the degree of color.
- Volumetric characterization.
- Relative tissue characterization.

Cell features consists of twenty-six descriptions. These properties were all extracted from cells, as described previously [35].

Normalization techniques used to unify the data description did not influence the classifier result.

The classification was conducted using two different techniques, Neural Network and Decision Tree. The Neural Network model yielded a better result, whereas the Decision Tree model was faster. We suggest using the Neural Network model based on its sensitivity to differentiate between each group of Acute Leukaemia cells.

The obtained results encourage future work to develop a robust segmentation system independent of stains used in blood smear images. The size of the dataset also needs to be expanded to provide the classification model with a greater number of useful examples in the training phase. This expansion should employ more Acute Myeloid Leukaemia cell types to differentiate all cells fully.

References

[1] S. Angelescu, N.M. Berbec, A. Colita, D. Barbu, A.R. Lupu, Value of multifaced approach diagnosis and classification of acute leukemias, Mædica 7 (3) (2012) 254–260.
[2] N.Y. Asaad, M. Dawoud, Diagnosis and prognosis of B-cell chronic lymphocytic leukemia/small lymphocytic lymphoma (B-CLL/SLL) and mantle cell lymphoma (MCL), J. Egypt. Nat. Cancer Inst. 17 (4) (2005) 279–290, December.
[3] A. Aimi Salihah, M.Y. Mashor, N.H. Harun, H. Rosline, Colour image enhancement techniques for acute leukemia blood cell morphological features, in: IEEE International Conference on Systems Man and Cybernetics, 2010, pp. 3677–3682.
[4] M.U.S. Daniela, d.F.C. Luciano, G.R. Edgar, A.Z. Marco, A texture Approach to Leukocyte recognition, Real Time Imag. 10 (4) (2004) 205–216, OCT.
[5] T. Zuva, O.O. Olugbara, S.O. Ojo, S.M. Ngwira, Image segmentation, available techniques, developments and open issues, Can. J. Image Process. Comp. Vis. 2 (3) (2011 March) 20–29.
[6] K. Kim, J. Jeon, W. Choi, P. Kim, Y.S. Ho, Automatic cell classification in human's peripheral blood images based on morphological image processing, in: AI 2001: Advances in Artificial Intelligence, Springer, Berlin Heidelberg, Australia, 2001, pp. 225–236.
[7] L.B. Dorini, R. Minetto, N.J. Leite, White blood cell segmentation using morphological operators and scale-space analysis, in: IEEE, XX Brazilian Symposium on Computer Graphics and Image Processing, 2007, Brazil, pp. 294–304.
[8] A.S. Prasad, K.S. Latha, S.K. Rao, Separation and counting of blood cells using geometrical features and distance transformed watershed, Int. J. Eng. Innov. Technol. (IJEIT) 3 (2) (2013).
[9] M. Madhukar, S. Agaian, A.T. Chronopoulos, Deterministic model for Acute Myelogenous Leukemia classification, in: IEEE International Conference on Systems Man and Cybernetics (SMC), 2012, pp. 433–438.
[10] Q. Chen, X. Yang, E.M. Petriu, Watershed segmentation for binary images with different distance transforms, in: Proceedings of the 3rd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, HAVE 2004, Ottawa, Ontario, Canada, 2004, pp. 111–116.
[11] H. Kekre, T.K. Sarode, Multilevel vector quantization method for codebook generation, Int. J. Eng. Res. Indust. Appl. (IJERIA) 2 (5) (2009) 217–231.
[12] H. Kekre, B. Archana, H.R. Galiyal, Segmentation of blast using vector quantization technique, Int. J. Comp. Appl. 4 (5) (2013).
[13] R.D. Labati, V. Piuri, F. Scotti, The acute lymphoblastic leukemia image database for image processing, in: 18th IEEE international conference on Image processing (ICIP), 2011; Università degli Studi di Milano, Department of Information Technology, via Bramante65, 26013 Crema, Italy, 2012, pp. 2045–2048.
[14] A. Korzynska, L. Roszkowiak, C. Lopez, R. Bosch, L. Witkowski, M. Lejeune, Validation of various adaptive threshold methods of segmentation applied to follicular lymphoma digital images stained with 3,3′-Diaminobenzidine&Haem, Diagn. Pathol. 8 (48) (2013).
[15] H. Hengena, S.L. Spoor, M.C. Pandi, Analysis of blood and bone marrow smears using digital image processing techniques, in: Medical Imaging 2002, vol. 4684, International Society for Optics and Photonics, 2002, pp. 624–635.
[16] P. Yan, X. Zhou, M. Shah, S.T. Wong, Automatic segmentation of high-throughput RNAi fluorescent cellular images, IEEE Trans. Inf. Technol. Biomed. 12 (1) (2008) 109–117.
[17] T. Schäfer, H. Schäfer, A. Schmitz, C. Döring, J. Ackermann, N. Dichter, et al, Image database analysis of Hodgkin lymphoma, Comput. Biol. Chem. 46 (September) (2013) 1–7.
[18] M. Mohamed, B. Far, An enhanced threshold based technique for white blood cells nuclei automatic segmentation, in: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), October 2012, Calgary, Canada, pp. 202–207.
[19] E. Mohammed, M.M. Mohamed, C. Naugler, B.H. Far, Chronic lymphocytic leukemia cell segmentation from microscopic blood images using watershed algorithm and optimal thresholding, in: 26th IEEE Canadian Conference of Electrical And Computer Engineering (CCECE), 2013, Canada.
[20] M. Madhukar, S. Again, A.T. Chronopoulos, New decision support tool for acute lymphoblastic leukemia classification, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2012, pp. 829518-829518–12.
[21] D.M.U. Sabino, Costa L. da Fontoura, E.G. Rizzatti, M.A. Zago, A texture approach to Leukocyte recognition, Real Time Imag. 10 (4) (2004) 205–216, OCT.
[22] F. Sadeghian, Z. Seman, A.R. Ramli, B.A. Kahar, M.I. Saripan, A framework for white blood cell segmentation in microscopic blood images using digital image processing, in: Shulin Li (Ed.), Biological Procedures Online, vol. 11(1), 2009.
[23] A. Belsare, M. Mushrif, Histopathological image analysis using image processing techniques: an overview, Sig. Image Process.: Int. J. (SIPIJ) 3 (4) (2012 August) 23.
[24] D.S. Kumar, G. Sathyadevi, S. Sivanesh, Decision support system for medical diagnosis using data mining, Int. J. Comp. Sci. Iss. 8 (3) (2011) 147–153.
[25] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comp. Appl. 17 (8) (2011).
[26] H.A. Abbass, An evolutionary artificial neural networks approach for breast cancer diagnosis, Artif. Intell. Med. 25 (3) (2002) 265–281.
[27] D.C. Huang, K.D. Hung, Y.K. Chan, A computer assisted method for leukocyte nucleus segmentation and recognition in blood smear images, J. Syst. Softw. 85 (9) (2012) 2104–2118.
[28] B. Pandey, R. Mishra, Knowledge and intelligent computing system in medicine, Comput. Biol. Med. 39 (3) (2009) 215–230.

[29] P.J. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, Neur. Netw. 15 (1) (2002) 11–39.
[30] M. Galindo, Obtaining Features of Subtypes of Leukemia in Digital Blood Cell Images for their Classification, Master Thesis, INAOE, Mexico, 2008.
[31] H.J. Escalante, M. Montes-y-Gómez, J.A. González, P. Gómez-Gil, L. Altamirano, C.A. Reyes, et al, Acute leukemia classification by ensemble particle swarm model selection, Artif. Intell. Med. 55 (3) (2012).
[32] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al, Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.
[34] H. Kekre, B. Archana, H.R. Galiyal, Segmentation of blast using vector quantization technique, Int. J. Comp. Appl. 72 (15) (2013).
[35] C. Reta, L.A. Robles, J.A. Gonzalez, R. Diaz, J.S. Guichard, Segmentation of bone marrow cell images for morphological classification of acute leukemia, in: The Twenty-Third International Florida Artificial Intelligent Research Society Conference FLAIRS Conference, 2010 May, Florida, pp. 86–91.
[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, ACM SIGKDD Expl. Newslett. 11 (1) (2009) 10–18.
[37] S.B. Gelfand, C. Ravishankar, E. Delp, An iterative growing and pruning algorithm for classification tree design, in: Conference on the Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 1989, 1991 February, pp. 163–174.

Further reading

[33] Lotufo ERDaRA, Hands-on Morphological Image Processing, SPIE, 2003.