Classifying acanthocytes using image processing and ML techniques: A comparative study

Catarina Silva[1]
c.alexandracorreia@ua.pt

Augusto Silva[1,2]
augusto.silva@ua.pt

Joaquim Madeira[1,2]
jmadeira@ua.pt

[1] Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro

[2] Institute of Electronics and Informatics Engineering of Aveiro, Universidade de Aveiro

Abstract

The diagnosis of several diseases can be improved with the identification of acanthocytes, i.e., red blood cells with abnormal form. We propose an approach to autonomously identify such cells in blood sample images. Our method relies on image processing operations and conventional machine learning methods. The principal motivation is the fact that this identification is usually performed by specialized devices or done manually by humans. Specialized devices are rare and costly, while manual identification is prune to error. Our approach reaches a precision of 91%, showing the potential of the solution.

1 Introduction

Red Blood Cells (RBC) are the most common cells present in the human body [5]. Normal RBC usually have a biconcave disk shape. If there is any abnormality in the shape of RBC, then it may indicate the presence of a disease. Furthermore, the shape and number of anomalous cells may also be an important indicator for medical diagnosis, improving its accuracy. It is important to segment and classify anomalous blood cells in order to detect diseases in a early stage, increasing the chances of successful recovery [1].

There are several types of anomalous blood cells, however we have focused our efforts on acanthocytes. The manual classification of abnormal cells under the microscope tends to give inaccurate results and errors [1]. Autonomous systems to detect and classify abnormal cells reduce the time needed to accomplish such task [3]. Furthermore, the latter typically have a lower error rate when compared to humans for that kind of repetitive work.

Our main objective is to develop a reliable detection and classification procedure for acanthocytes, using a reduced set of features. Image processing techniques are used to segment blood cells and conventional Machine Learning (ML) models to classify them. The output is the classification of each blood cell into one of two classes: normal cells or acanthocytes. Additionally, the number of acanthocytes in the blood sample is computed.

This paper is organised as follows: Section 2 presents the relevant background; in Section 3 the stages of the proposed approach are described; some details regarding the implementation are presented in Section 4; results are presented in Section 5; Section 6 presents some conclusions and ideas for future work.

2 Background

RBC suffer anomalies related with shape, size and color. According to [7], acanthocytes are "Erythrocytes with a dented and prickly profile with spicules of different lengths". The presence of acanthocytes is a strong indicator of several diseases, such as alcoholic cirrhosis, neonatal hepatitis and poor absorption states.

Several authors have developed methods to autonomously detect the presence of acanthocytes in medical images [4]. Two recent works [7, 8] proposed solutions based on image processing and classification methods.

The first work [7] applies morphological operations to extract the contour of the RBC and computes several features related with contour shape, such as: chain code, circularity and skeleton. After that, they use k-NN [9] as a classification algorithm to classify the extracted contours.

The second work [8] relies on image segmentation as a method to correctly extract the region of each individual blood cell. They also use ML methods for the final classification, namely k-NN and SVM [9].

3 Proposed approach

The proposed approach is based on [7] with some key differences. The image processing workflow is enhanced with the goal of improving region segmentation and contour extraction. We consider a reduced set of features that still contains enough information to properly classify the RBC while speeding-up the ML training time. Finally, we evaluated several ML models instead of only relying on k-NN.

The image processing pipeline is composed by several stages with the objective of reducing noise, enhancing region contours and segmenting them. The key stages of the pipeline are depicted in Figure 1.

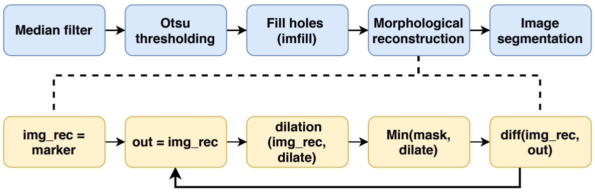
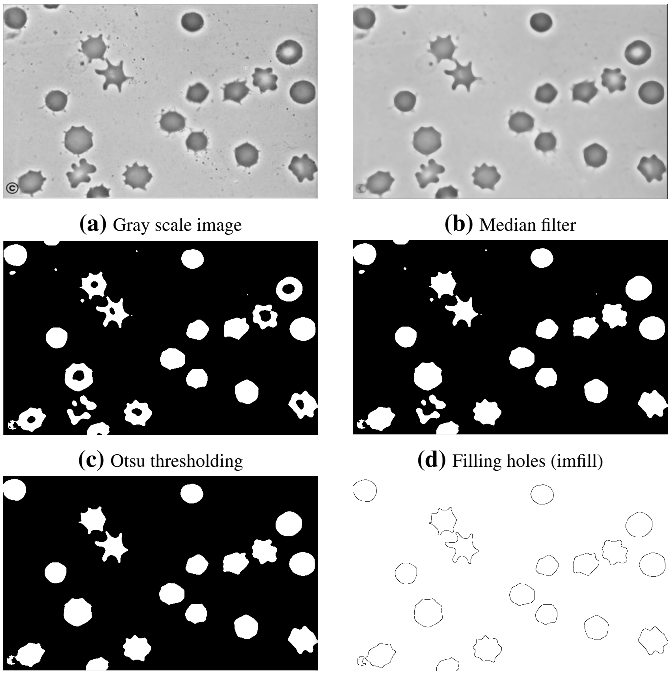


Figure 1: Image processing pipeline.

The first step is to normalize an input image by converting it to gray scale and applying a 9x9 median filter to smooth noise. The gray image is then converted to binary using the Otsu thresholding method. Those operations may originate some holes in the middle of the cells and medium-sized noise (by-product of the binarization).

The next steps fix that by executing a filling operation (imfill) that applies a guided flooding operation to close holes inside blobs. Morphological reconstruction (elliptic shaped 9x9 kernel) is applied to remove the medium-sized noise produced during the binarization. Finally, the Canny edge detector is applied to extract region contours. Figure 2 illustrates the results of the image processing pipeline.


(a) Gray scale image

(b) Median filter

(c) Otsu thresholding

(d) Filling holes (imfill)

(e) Morphological reconstruction

(f) Image segmentation (negative)

Figure 2: Result of the image processing pipeline.

Some care is needed when selecting features: since we are dealing with images, it is important to compute features that correctly express shape characteristics but are also invariant to rotation and scaling [2].

Based on the extracted region contours, we compute several features that describe them. The first feature is the histogram from the chain code. A chain code characterizes the shape of a contour but is not rotation and scale invariant. To achieve that we compute an histogram with the relative weight for each direction of the chain code. The remaining features are circularity, roundness, aspect-ratio and solidity. The previously mentioned features are shape descriptors commonly used by image processing toolboxes to classify blobs. These features are meant to enhance the classification process, by expanding the expressiveness of the histogram, and capture characteristics that are invariant to scale and rotation.

The chain code histogram, captures the complete shape into a condensed form. The remaining features capture the smoothness and stability of the shape. A smooth round blood cell will have a high circularity value, an aspect-ratio close to one and a high solidity value. The irregular shape of an acanthocyte will diverge from the previously mentioned values.

4 Implementation

The project code was developed in C++ using the Open Source Computer Vision (OpenCV) library[1] and it is hosted on the GitHub public repository[2].

All the image processing and feature extraction code was fully developed. It is important to mention that the imfill[3] and morphological reconstruction operations do not exist in OpenCV and were implemented. Furthermore, OpenCV offers limited support regarding chain codes, we implemented a method to follow each of the contour pixels and build a chain code sequence. Two classifiers were also implemented: k-NN and Logistic Regression.

The current prototype is composed by two main programs: one uses a dataset to train the previously mentioned models, the second uses the model to classify and count acanthocytes in blood cell images. In order to evaluate our prototype using other ML methods we implemented a method that outputs the features into a ARFF[4] dataset. An ARFF dataset can then be loaded into WEKA [6], a popular ML framework.

5 Results

We built a dataset for the evaluation of our prototype, by gathering medical images from microscopic blood samples. Based on the definition of acanthocytes, and on input from medical professionals, we manually segmented those images and classified each blood cell into the healthy and acanthocyte classes. The dataset is composed of 140 segmented images, 72 samples for the acanthocyte class and the remaining 68 samples as the healthy blood cells. The segmented images were resized to have a height of 96 pixels while maintaining the aspect ratio. The dataset is publicly available and can be found on the repository alongside the code.

The segmented images are processed by our proposed image processing pipeline, generating a ARFF output. After that we used the WEKA framework to explored and evaluate several ML models (not being limited by the kNN and logistic regression previsouly developed). It is important to mention that the default hyper-parameters were used for each model.

Three different metrics were used to evaluate the performance of the models: Precision, F-Measure and Matthews correlation coefficient (MCC). The models were evaluated with 10-fold cross validation. The results can be found on Table 1

All models achieve close to 70% precision demonstrating the potential of the developed approach. The top three models are: Random Forest, Neural Network (multi-layer perception) and Decision Tree. One interesting aspect regarding the decision tree model is that the model only uses 5 features: solidity, circularity, aspect ratio, h5 and h3 (h0 to h7 are the values from the chain code histogram). In other words, the model selects as relevant features solidity and circularity, while the remaining three (aspect ratio, h5 and h3) are used only for corner cases. Furthermore, it does not rely on the full histogram for the classification and only uses two of the eight available values.

Table 1: Classifier performance

ML Algorithm	Precision	F-Measure	MCC
k-NN(1)	0.710	0.704	0.415
k-NN(3)	0.709	0.684	0.400
k-NN(5)	0.748	0.723	0.476
Naive Bayes	0.680	0.652	0.342
Logistic Regression	0.867	0.864	0.731
Decision Tree	0.879	0.879	0.757
Random Forest	**0.910**	**0.909**	**0.819**
Support Vector Machine	0.711	0.630	0.363
Neural Network	0.886	0.886	0.773

6 Conclusions

We proposed a new approach for acanthocyte detection and classification based on image processing and ML models. Contrary to the current trend, we achieved a high precision without relying on Deep Neural Networks, that require substantial amount of data and time to train effectively.

The proposed prototype achieves 91% precision, demonstrating the potential of the solution. We intend to improve our prototype by testing other image segmentation methods (*e.g.* watershed) and convert the code into Python to leverage the advanced ML frameworks available. Furthermore, we intend to explore the importance of each feature that composes our dataset and devise new ones that can improve the accuracy.

References

[1] H. A. Aliyu, R. Sudirman, M. A. Abdul Razak, and M. A. Abd Wahab. Red blood cell classification: Deep learning architecture versus support vector machine. In *2nd Int Conf on BioSignal Analysis, Processing and Systems (ICBAPS)*, pages 142–147, 2018.

[2] Mário Antunes and Luís Seabra Lopes. Unsupervised internet-based category learning for object recognition. In *Lecture Notes in Computer Science*, pages 766–773. Springer, 2013.

[3] S. F. Bikhet, A. M. Darwish, H. A. Tolba, and S. I. Shaheen. Segmentation and classification of white blood cells. In *IEEE Int Conf on Acoustics, Speech, and Signal Processing*, volume 4, pages 2259–2261, 2000.

[4] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12 – 22, 2019.

[5] P. T. Dalvi and N. Vernekar. Computer aided detection of abnormal red blood cells. In *IEEE Int Conf on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 1741–1746, 2016.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, nov 2009.

[7] María Elena Cruz Meza, MengYi En, Graciela Vázquez Álvarez, and José Cruz Martínez Perales. Detection and classification of abnormalities in erythrocytes by techniques of image analysis and pattern recognition. In *16th LACCEI International Multi-Conference for Engineering, Education, and Technology*, pages 18–20, July 2018.

[8] Cecilia Di Ruberto, Andrea Loddo, and Lorenzo Putzu. A region proposal approach for cells detection and counting from microscopic blood images. In *Lecture Notes in Computer Science*, pages 47–58. Springer, 2019.

[9] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification. *International Journal of Data Warehousing and Mining*, 3(3):1–13, jul 2007.

[1] https://opencv.org/

[2] https://github.com/catarinaacsilva/medical-image-processing

[3] https://www.mathworks.com/help/images/ref/imfill.html

[4] https://www.cs.waikato.ac.nz/ ml/weka/arff.html