Check for updates

RESEARCH ARTICLE

Sickle cell segmentation and classification for thalassemia aid diagnosis [version 1; peer review: 1 approved]

Yen-Siang Leow¹, Kok-Why Ng ¹, Yih-Jian Yoong¹, Seng-Beng Ng²¹Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia²Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), UPM Serdang, Selangor, 43400, Malaysia

V1 First published: 23 Nov 2021, 10:1185
<https://doi.org/10.12688/f1000research.73314.1>
Latest published: 23 Nov 2021, 10:1185
<https://doi.org/10.12688/f1000research.73314.1>

Abstract

Background: Thalassemia is a hereditary blood disease in which abnormal red blood cells (RBCs) carry insufficient oxygen throughout the body. Conventional methods of thalassemia detection through a complete blood count (CBC) test and peripheral blood smear image still possess a lot of weaknesses.

Methods: This paper proposes a hybrid segmentation method to segment the RBCs. It incorporates adaptive thresholding and canny edge method to segment the RBCs. Morphological operations are performed to clean the leftovers. Shape and texture features are extracted using the segmented masks and the gray level co-occurrence matrix. Data imbalance treatment is used for solving the imbalance cell type class in distribution. In the data resampling layer, the synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling (ADASYN), and random over sampling (ROS) are performed and evaluated using the decision tree and logistic regression. In the classification layer, the decision tree, random forest classifier and support vector machine (SVM) are assessed and compared for the best performance in classification.

Results: The proposed method outperforms the other methods in the image segmentation layer with the structural similarity index measure (SSIM) of 89.88%. In the data resampling layer, ADASYN is employed as it is more accurate than the SMOTE and ROS. The random forest classifier is chosen at the classification layer as it is more accurate than the decision tree and support vector machine (SVM).

Conclusions: The proposed method is tested on the latest dataset of erythrocyteIDB3 and it solves the issues of imbalanced data due to the insufficient cell classes.

Keywords

Red Blood Cell, Thalassemia, Image processing, Image Segmentation, Morphological Operations, Feature Extraction, Data Resampling, Machine Learning

Open Peer Review

Approval Status

1

version 1

23 Nov 2021

[view](#)

1. Naresh Babu Muppalaneni , National Institute of Technology Silchar, Silchar, India

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research Synergy Foundation gateway**.

This article is included in the **Sickle Cell Disease collection**.

Corresponding author: Kok-Why Ng (kwng@mmu.edu.my)

Author roles: **Leow YS:** Conceptualization, Formal Analysis, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Ng KW:** Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Review & Editing; **Yoong YJ:** Conceptualization, Supervision, Writing – Review & Editing; **Ng SB:** Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is supported by the funding of Multimedia University IR Fund (Ref: MMUI/210005).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Leow YS *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Leow YS, Ng KW, Yoong YJ and Ng SB. **Sickle cell segmentation and classification for thalassemia aid diagnosis [version 1; peer review: 1 approved]** F1000Research 2021, **10**:1185 <https://doi.org/10.12688/f1000research.73314.1>

First published: 23 Nov 2021, **10**:1185 <https://doi.org/10.12688/f1000research.73314.1>

Introduction

Sickle cell disease is a hereditary disease that disrupts the efficiency of RBCs to carry sufficient oxygen to the body. Traditional methods to detect this disease relied on hematologists manually counting RBCs and classifying them based on their topologies. This has motivated us to explore and develop an automated process to detect the presence of the disease. However, this field poses a lot of challenges such as insufficient medical data, poor extraction methods, and the tedious pipeline-related issues.

This paper aims to solve some of the aforementioned problems. Firstly, a hybrid segmentation method is presented to accurately segment the RBCs. Next, the data imbalance treatment is used to solve the imbalance cell type class in a distribution manner. The pipeline of the paper includes the comparison between the conventional methods in the image segmentation layer, data resampling layer and the classification layer. The proposed method involves the image acquisition, image pre-processing, image segmentation, morphological operations, image cropping and manual classification, feature extraction, data resampling, and classification. The aim is to create a workflow that is able to detect the presence of thalassemia by recognizing the types of RBCs.

Related work

In image pre-processing, Rashid *et al.*¹ and Das *et al.*² employed a mixture of global contrast techniques, median blurring and color space conversion to increase the contrast. Color channel conversion is used to filter the unwanted cells in the image.

In image segmentation layer, an inter-combination of the thresholding, edge-based, region-based, clustering-based and artificial neural network (ANN) is used to perform the segmentation. Sharma *et al.*³ compared different techniques of thresholding; their findings showed that Otsu's thresholding performed the best among other thresholding methods. This concurred with the observation done by Das *et al.*² Marked water controlled segmentation was used to solve the segmentation of cells in.^{2,4} Tyagi *et al.*⁵ indicated that the Canny Edge technique was better at filtering noise presented in the image. Achargee *et al.*⁶ used Hough transformation (HT) to detect cells with varying diameter, while Seth and Palodhi⁷ proposed a modified version of the HT to detect white blood cells (WBC).

Feature extraction in cell morphology are classified into boundary and region descriptors. Some of the features may overlap, but generally, boundary descriptors deal with the morphology of the RBCs and region descriptors extract the geometric and texture properties. Lotfi *et al.*⁴ used Fourier descriptors to represent cell morphology. Tyagi *et al.*⁵ used shape features such as area, perimeter solidity, eccentricity to describe the cells. Other features involved extracting the moment invariant such as invariant to scaling, translation and rotation were proposed by Lotfi *et al.*⁴ Since the introduction of deep learning, many skipped this layer due to the models can automatically extract the features. Fadhel *et al.*⁸ proposed extracting features from the last fully connected layers and subsequently fed the features into the support vector machine (SVM) classifier. Purwar *et al.*⁹ proposed a hybrid method of features extracted from the convolutional neural network (CNN) models¹⁵⁻¹⁷ and hand clinical features extracted from a CBC test.

Hortinela *et al.*¹⁰ and Safca *et al.*¹¹ used SVM to classify cells based on features extracted from different layers. In multiclass classification, Lotfi *et al.*⁴ compared the K-nearest neighbors and SVM method using the one *versus* all approach and found that SVM performed better. In a separate study, Dalvi *et al.*¹² showed that an ANN with ten hidden nodes and five output nodes performed better than decision tree. Tyas *et al.*¹³ used ANN with momentum backpropagation and produced the best results among the classifiers.

Methods

Ethical approval number: EA1542021

Ethical approval body: Research Ethic Committee 2021, Multimedia University

Image acquisition

The dataset used in this project was obtained from the erythrocytesIDB by Pedro *et al.*¹⁴ It is selected because of the image clarity and optimal cell size. Each subdirectory contains the raw image, ground truth labels, and binary segmented masks. Figure 1 shows the raw blood smear images and its ground truth label for two images.

Image pre-processing

As many of the images are extracted using different extraction techniques, to increase the quality of the images while standardizing all of the images, firstly, the images are contrastly increased by using the contrast limited adaptive histogram equalization (CLAHE) method. The image is primarily converted to L.A.B. color space. The L, which refers to the luminance channel is operated by CLAHE and is then merged back with the other two channels.

Figure 1. Raw image and ground truth label.

Next, median filtering is applied to remove noise in the image. An unsharp mask operation is applied to sharpen the edges of the cells which are previously lost during the median filtering process. Image subtraction is then constructed from the original image to obtain the details and the edges. **Figure 2** depicts the result after the pre-processed layer of the two images.

Figure 2. Image before and after pre-processing.

Image segmentation

A hybrid technique is proposed by merging the adaptive thresholding and the canny edge detection method. Gaussian adaptive thresholding is employed to segment the color images with varying intensities. However, this method unable to capture the edges of the cells. An additional of the canny edge detection method with the parameter of μ to be set to a lower value is applied. This allows the cell edges to be captured precisely and subsequently to be added to the threshold binary image later.

Lastly, the leftover edges that are protruding out are removed to show a clean finish. **Figure 3** shows the result of the segmentation layer compared to the ground truth given from the dataset. The evaluation of this layer is performed by calculating the SSIM index of the binary segmented image with the ground truth binary given in the dataset.

Morphological operations

Morphological operations are performed on binary images to clean up the leftover noises from the previous layers. A hole-filling operation is performed to close up the holes that are created using the canny edge detection method. Next, binary opening is used to remove the small particles that are presented in the image. Finally, cells that are located at the border are removed because they do not accurately provide the full feature description of the cell. **Figure 4** depicts the result of the morphological operations.

Image cropping and classification

Each of the cells in the microscopic blood sample images are cropped using the masks that are segmented from the previous layers. **Figure 5** shows the individual cells that are being cropped, while **Figure 6** depicts the sub-category of one particular cell. Each of the cropped cells has their own directory which is named based on their origin and cell number as depicted in **Figure 7**. The cells are manually classified into their own class which is based on the classification given by Pedro *et al.*¹⁴

Feature extraction

Two types of features are extracted: shape features and texture features.

Figure 3. Segmented image to be compared with the ground truth mask.

Figure 4. Image after morphological operations to be compared with the ground truth mask.

Figure 5. Cells that are being cropped.

My Drive > ... > Circular > 30erythrocytesIDB3_19 ▾

Figure 6. Sub-directory of one cell.

Figure 7. Directory of the circular cells.

The shape features are extracted from the individually cropped masks and then to be combined into a data frame. The features extracted include the area, perimeter, major axis length, minor axis length, eccentricity, and the seven moment invariants immune to the scaling, translation, and rotation. Additional shape features which are circulatory factor and elliptical factor are calculated.

Next, texture features are extracted using the gray level co-occurrence matrix (GLCM). The features extracted from the textures are the contrast, dissimilarity, homogeneity, ASM, energy and correlation.

Finally, the two features are merged into a data frame with its correct label of either one of the three classes: circular, elongated, or other cells.

Data resampling

The number of circulars, elongated and other types of cells total up to 316, 53 and 59 respectively. This dataset is imbalanced and may result the classifier to be overfit or bias to certain category of data.

Figure 8. ADASYN data resampling.

The synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling (ADASYN) and random oversampling (ROS) are evaluated by using two classifiers, which are logistic regression and decision tree. Figure 8 shows the number of classes prior and after the resampling process. The bottom graph of Figure 8 demonstrates the best result for solving the imbalanced problem using ADASYN resampling technique.

Classification

In classification, a total of three classifiers are trained and compared. The chosen classifiers are the decision tree, random forest and the support vector machine.

The models are evaluated using the classification report which contains the information of accuracy, precision, recall f1-score and support.

For each classifier, the receiver operating characteristic (ROC) curve is plotted to observe the results. The images that are excluded before the sampling process are tested individually for their accuracy and reliability of the model chosen.

Results and discussion

Image segmentation layer

The ground truth masks are provided in the dataset. The evaluation are calculated based on the structural similarity index (SSIM) between the ground truth masks and the segmentation of the proposed method.

The average of the computed SSIM index from each image is 89.88%. However, it is worth noting that the value only estimates how well the segmentation is. In [Figure 9](#), observing the ground truth mask, some cells are not presented in the ground truth mask (middle image), while it is presented in the proposed segmented image (right most image).

Data resampling layer

[Figure 10](#) shows the comparison of the resampling techniques between the SMOTE, ADASYN and ROS. Based on the result, logistic regression handles the unbalanced data-sets well as compared with the decision tree. ADASYN outperforms the rest at 84.49% of its accuracy through the logistic regression classifier. Also worth noting is that, even with the unbalanced dataset, its accuracy sits highly at 82.17%, which indicates that the features extracted optimally best describe the types of cells ([Figure 11](#)).

Figure 9. Cells missing in ground truth mask.

Algorithm	Method	Accuracy	F
Logistic Regression	Unbalanced	0.821705	
Logistic Regression	SMOTE	0.821705	
Logistic Regression	ADASYN	0.844961	
Logistic Regression	ROS	0.821705	
Decision Tree	Unbalanced	0.798450	
Decision Tree	SMOTE	0.744186	
Decision Tree	ADASYN	0.767442	
Decision Tree	ROS	0.790698	

Figure 10. Comparison of different sampling techniques.

Classification

The classification layer is evaluated based on the accuracy of the classification report and the ROC curve of each classifier. Figures 12 to 14 depict the model evaluation summary for each of the classifiers.

Figure 11. Resampling techniques evaluation bar plot.

Figure 12. Decision tree model evaluation.

Figure 13. Random forest model evaluation.

Figure 14. SVM model evaluation.

From the results, the random forest classifier performs the best with the accuracy of 90% while the others are below 90%.

The ROC curve for the SVM dips below the threshold where the number of incorrectly classified classes are more than the correctly classified classes, which indicates a bad classifier.

Comparing the area under curve (AUC) of the ROC curve, the random forest classifier covers the most area for all the three cell classes of the circular, elongated and other types of cell, which demonstrates a superior classifier.

Conclusion

We performed several comparisons and the implementation between different techniques in the segmentation layer, data resampling layer and the classification layer. These provide a good working pipeline for a complete aiding diagnosis for the sickle cell disease detection. The whole pipeline provides a precise classification for each of the cells in microscopic blood smeared images with a 90–95% of accuracy on average.

In addition, we present a strong baseline in the segmentation layer, which is able to detect some of the missing cells that were not presented in the previous work. At the same time, the data resampling layers are added to observe the effect of an unbalanced medical dataset. The combination of different models exhibits the performance of the best classification model.

Author contributions

Leow YS, Ng KW, Yoong YJ and Ng SB conceived the presented idea. Leow YS carried out the experiment and wrote the manuscript. Ng KW, Yoong YJ and Ng SB supervised the project and provided feedback.

Data availability

Underlying data

Data were obtained by permission from <http://erythrocytesidb.uib.es/>.

Data samples were obtained from volunteer patients. This dataset was not generated nor is it owned by the authors of this article; the listed owners are Universidad de Oriente, Cuba, and Universitat de les Illes Balears, Spain. Therefore, neither the authors nor F1000Research are responsible for the content of this dataset and cannot provide information about data collection.

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments and suggestions.

References

1. Rashid NZN, Mashor MY, Hassan R: **Unsupervised color image segmentation of red blood cell for thalassemia disease.** 2015 2nd International Conference on Biomedical Engineering (ICoBE). IEEE; 2015, March; (pp. 1–6).

2. Das D, Ghosh M, Chakraborty C, *et al.*: **Invariant moment based feature analysis for abnormal erythrocyte recognition.** 2010 *International Conference on Systems in Medicine and Biology*. IEEE; 2010, December; (pp. 242–247).
3. Sharma V, Rathore A, Vyas G: **Detection of sickle cell anaemia and thalassaemia causing abnormalities in thin smear of human blood sample using image processing.** 2016 *International Conference on Inventive Computation Technologies (ICICT)*. IEEE; 2016, August; (Vol. 3: pp. 1–5).
4. Lotfi M, Nazari B, Sadri S, *et al.*: **The detection of dacrococyte, schistocyte and elliptocyte cells in iron deficiency anemia.** 2015 *2nd International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE; 2015, March; (pp. 1–5).
5. Tyagi M, Saini LM, Dahyia N: **Detection of poikilocyte cells in iron deficiency anaemia using artificial neural network.** 2016 *International Conference on Computation of Power, Energy Information and Commuincation (ICCPDIC)*. IEEE; 2016, April; (pp. 108–112).
6. Acharjee S, Chakraborty S, Alam MI, *et al.*: **A semiautomated approach using GUI for the detection of red blood cells.** 2016 *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE; 2016, March; (pp. 525–529).
7. Seth S, Palodhi K: **An efficient algorithm for segregation of white and red blood cells based on modified hough transform.** 2017 *IEEE Calcutta Conference (CALCON)*. IEEE; 2017, December; (pp. 465–468).
8. Alzubaidi L, Fadhel MA, Al-Shamma O, *et al.*: **Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis.** *Electronics*. 2020; 9(3): 427.
[Publisher Full Text](#)
9. Purwar S, Tripathi RK, Ranjan R, *et al.*: **Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers.** *Multimed. Tools Appl*. 2020; 79(7): 4573–4595.
[Publisher Full Text](#)
10. Hortinela CC, Balbin JR, Fausto JC, *et al.*: **Identification of Abnormal Red Blood Cells and Diagnosing Specific Types of Anemia Using Image Processing and Support Vector Machine.** 2019 *IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. IEEE; (pp. 1–6).
11. Safca N, Popescu D, Ichim L, *et al.*: **Image processing techniques to identify red blood cells.** 2018 *22nd International Conference on System Theory, Control and Computing (ICSTCC)*. IEEE; 2018, October; (pp. 93–98).
12. Dalvi PT, Vernekar N: **Computer aided detection of abnormal red blood cells.** 2016 *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE; 2016, May; (pp. 1741–1746).
13. Tyas DA, Ratnaningsih T, Harjoko A, *et al.*: **The classification of abnormal red blood cell on the minor thalassemia case using artificial neural network and convolutional neural network.** *Proceedings of the International Conference on Video and Image Processing 2017*, December; (pp. 228–233).
14. Pedro MF, Grethel CS, Wilkie EDF, *et al.*: **erythrocytesIDB (Version 2, October 2017).**
[Reference Source](#)
15. Acharya V, Kumar P: **Identification and red blood cell classification using computer aided system to diagnose blood disorders.** 2017 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE; 2017, September; (pp. 2098–2104).
16. Ong K, Haw SC, Ng KW: **Deep Learning Based-Recommendation System: An Overview on Models, Datasets, Evaluation Metrics, and Future Trends.** *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems 2019*, November; (pp. 6–11).
17. Ang JS, Ng KW, Chua FF: **Modeling Time Series Data with Deep Learning: A Review, Analysis, Evaluation and Future Trend.** 2020 *8th International Conference on Information Technology and Multimedia (ICIMU)*. IEEE; 2020, August; (pp. 32–37).

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 09 February 2022

<https://doi.org/10.5256/f1000research.76957.r100976>

© 2022 Muppalaneni N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Naresh Babu Muppalaneni

National Institute of Technology Silchar, Silchar, Assam, India

The authors have done extensive and novel work to bring this scientific work.

For Sickle cell, the circulatory feature is important, which is considered by authors in the feature set. The authors have extracted features based on shape and texture. The authors have addressed the imbalance problem also.

Please include the following in the next version of the paper:

1. Comparative studies of authors proposed methods and with literature
2. Missing dataset details like the size of the dataset, training set size, test dataset size, etc.,
3. Reasons for the vast difference in accuracy for SVM and Decision Tree/Random Forest.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, Deep Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research