Plasmodium classification on red blood cells image using multiclass support vector machines

S F Y O Pradini*, A Bustamam¹, and Z Rustam¹

¹Department of Mathematics, FMIPA, Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

*Corresponding Author: soya.febeauty@gmail.com

Abstract. Classification methods have been frequently used in various aspects, including bioinformatics. One of it’s purpose of this classification is to determine phase level of a disease. This research will classify the phase of plasmodium falciparum parasite which causes malaria. The disease is spread by an infected female Anopheles mosquito, which contains Plasmodium. The result of this research could be used to determine Plasmodium parasite phase in infected people’s red blood cells. The purpose of this research is to discover the success rate of Multiclass Support Vector Machines method and analyze it in order to predict the parasite phase levels. The data of this study is image data of red blood cells which were infected by three kinds of Plasmodium falciparum parasite levels. In the process, this study will be using Canopy as Integration Development Environments of phyton programming language. From 112 trials, the highest number of accuracy is 87.5% for Multiclass Support Vector Machines one vs rest method which used the 4-fold cross-validation with C=1 as parameter for linear kernel.

1. Introduction

Plasmodium is a single-celled of the class of Sporozoa that includes the parasite that causes malaria. This disease is spread by an infective female Anopheles mosquito which contains Plasmodium that will breed in human red blood cells [1].

In World Malaria Report 2015, at least 106 countries have been infected by malaria, including Indonesia. Based on the data collected from Direktorat Jenderal Pencegahan dan Pengendalian Penyakit Kementerian Kesehatan Republik Indonesia in 2016, Indonesia already has a good treatment to prevent the transmission of malaria and there is an enhancement of blood availability check percentage gradually[2]. In 2015 the percentage of blood availability check target had reaching 99% while the previous target was 95%. It shows that laboratory blood tests in almost every case of malaria suspect had been done [3].

Research about malaria in Indonesia has shown that the disease mostly caused by Plasmodium falciparum. In 2017, that parasite had been found in the bone marrow of a child in Manado City [4] and [5] showed that there are 62% of the malaria case in eastern Indonesia which inflicted by Plasmodium falciparum. As for the treatment, according to infoDatin [3] the cure will depend on the parasite’s class. Hence, this research will discuss about classification of Plasmodium falciparum parasite class in red blood cells. Research about that classification using digital image processing and Region Property computation method show the accuracy of 96% [6]. Based on that, this research used Multiclass Support Vector Machines (SVM) method to determine the class of Plasmodium falciparum parasite levels.
2. Method
For SVM classification, we need to calculate the variables for the feature. Then, we used Multiclass SVM with help of Canopy Software for data classification. The steps will be shown in Figure 1:

![Figure 1. Schematic Diagram for Classification Method using Multiclass SVM](image)

2.1. Data
The data consist of 63 red blood cells from 19 red blood cell images, which contains *Plasmodium falciparum* in many phase levels. It divided into four kinds of phase levels: normal, trophozoite, gametocyte, and schizont.

![Figure 2. Plasmodium falciparum phases in infected red blood cells: (a) trophozoite, (b) gametocyte, (c) schizont](image)

2.2. Variable
This research takes several feature variables that will be used for the classification process. The feature was taken from RGB color histogram from available blood cells image. Color histogram was obtained from normalized color intensity levels calculation. To determine the histogram, we will use these equations:

\[
\text{Normalized Red} = \frac{\text{Red}}{\sqrt{\text{Red}^2 + \text{Green}^2 + \text{Blue}^2}}
\]

\[
\text{Normalized Green} = \frac{\text{Green}}{\sqrt{\text{Red}^2 + \text{Green}^2 + \text{Blue}^2}}
\]

\[
\text{Normalized Blue} = \frac{\text{Blue}}{\sqrt{\text{Red}^2 + \text{Green}^2 + \text{Blue}^2}}
\]

Then, we will calculate the features. There are a total of 12 features that will be useful, which are: mean, variance, upper bound and lower bound for red, blue and green colors.
2.3. Thresholding
Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images. In image processing, it is used to split an image into smaller segments, or junk, using at least one color grayscale value to define their boundary [7]. Research about using thresholding to find infected red blood cells showed an accuracy of 92.85% [8]. And so, we will use 160 as the boundary. It means, if the pixel’s color is more than 160, then it will turn into 1; otherwise, it will be 0.

2.4. Support Vector Machines
SVM was first introduced to solve the pattern classification and regression problems by Vapnik and his colleagues [9]. This method uses a linear model to implement nonlinear class boundaries through some non-linear mapping input vectors into a high-dimensional feature space. The linear model constructed in the new space can represent a nonlinear decision boundary in the original space. In the new space, an optimal separating hyperplane is constructed. Thus, SVM is known as the algorithm that finds a special kind of linear model, the maximum margin hyperplane [10]. It is said that SVM training and operation requires low computational power, which gives potential for real-time operation [11].

2.4.1. Multiclass Method. SVM is a method that will classify not only binary class problems but also a multiclass classification problem. According to Liu, Bi, and Fan [12], one of the most commonly used methods in multiclass classification is one-vs-one (OVO) multiclass method. This method will be transforming multi-class problem into multiple binary classification problem. On it’s process, an n-class classification problem will be transforming into n(n-1)/2 binary classification problems. Hsu and Lin [13] gave an example to solve the binary classification problem for training data from ith and jth classes :

\[
\min_{w^i, b^i, \xi^i} \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_t \xi^i_t (w^{ij})^T
\]

\[
(w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi^i_t, \text{ in subject to } y_t = i
\]

\[
(w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi^i_t, \text{ in subject to } y_t \neq i
\]

Another method for multiclass classification is one-vs-rest (OVR). According to Permata, Ariwibowo and Maulana[14], it is a strategy to build an n-class binary and compare each data to the rest of sample. In their research, Hsu and Lind [13] state that the method will construct k SVM models where k is the number of classes. The ith class with positive labels and all other examples with negative labels. Thus given / training data (x₁, y₁), ... , (x_l, y_l), where xᵢ ∈ ℝⁿ, i = 1, ..., l and yᵢ ∈ {1, ..., k} is the class of xᵢ, the ith SVM solves the following problem :

\[
\min_{w^i, b^i, \xi^i} \frac{1}{2} (w^i)^T w^i + C \sum_{j=1}^l \xi^i_j (w^i)^T
\]

\[
(w^i)^T \phi(x_j) + b^i \geq 1 - \xi^i_j, \text{ in subject to } y_j = i
\]

\[
(w^i)^T \phi(x_j) + b^i \leq -1 + \xi^i_j, \text{ in subject to } y_j \neq i
\]

\[
\xi^i_j \geq 0, \ j = 1, ..., l
\]

3. Results and Discussion
In this study, we used the data from Centers of Disease Control and Prevention which can be accessed from it’s site [15]. In pre-processing, each image will be resized so we will have a 300x300 pixel as the dimension, so every image has the same size. Then, we do mask and thresholding for each of it so that
the object will be separate from the background. The result will be black and white images which the
color those values below 160 will turn to black and for the rest is white.

After getting the black and white images, we will create bounding box for every identified object.
Bounding box is needed so we could focus the analyzing process just on the area inside the bounding
box. Since there might be more than one white area on the images, there will be more than one bounding
box, hence we got 63 bounding boxes from 19 red blood cell images. When all of the images got their
bounding box, then we put the bounding box onto the color images.

![Figure 3. Bounding box from black and white images being put onto the color images](image)

For there are different kind of images, there will be different kind of objects inside the bounding box,
even the normal ones goes detected. So, we decided to add normal class, therefor, we had four classes
for the classification aside from *trophozoite*, *gametocyte* and *schizont*. Then, we will create color
composition histogram for every bounding boxes.

We will do classification using linear and RBF kernel with \{1, 10, 100, 1000\} as the value of \( C \) for
both kernel and \{0.0001, 0.001, 0.01, 0.1, 0.2, 0.5\} as the value of \( \gamma \) for RBF kernel, so that we have
112 classification results to compare in order to find the best parameter. This method shows that there
are some confusion matrices that has the same result.

For OVO classification, the highest percentage is 68.75% for 2-fold cross validation and 75% for 4-
fold cross validation. Meanwhile, OVR had 75% for 2-fold cross validation and 87.5% for 4-fold cross
validation. All of the highest score used linear kernel with \( C=1 \).

4. Conclusion

In this paper, we used Multiclass SVM as classification method with two types of kernel : linear and
RBF. The classification results show that we got the highest accuracy when we used linear kernel with
\( C=1 \). There is 68.75% accuracy for 2-fold cross validation for OVO and 75% accuracy for OVR. As for
4-fold cross validation, there is 75% accuracy for OVO and 87.5% for OVR.

These level of accuracy possibly because the lack of data, no feature selection process and different
types of background color for red blood cells images so that color intensity level had increase in number.
For further research, we propose to use more data with the same image’s color and also the different
types of features.

References

[1] Ashley E A, Phyo A P and Woodrow C J 2018 *The Lancet* **391** 1608
[2] IDI 2017 *Buku Saku Penatalaksanaan Kasus Malaria* (Directorate Generale of Disease
Prevention and Control: Jakarta)
[3] InfoDATIN 2016 *Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia*
ISSN :2442-7659
[4] Tatura S N N, Gunawan S and Bernadus 2017 *Asian Pasific Journal of Tropical Medicine* **10**
1015
[5] Hutagalung J 2017 *Lancet Glob. Health Spec. Issue* **5** S13
[6] Maspriyanti F, Nursari S R C, Murtako A and Gatc J 2016 *Information Systems and Electrical*
[7] Bhargavi K and Singaraju J 2014 *Int J. of Innovative Research & Dev. Special Issue* 3 234
[8] Gate J, Mapiyanti F, Sarwinda D and Arymurthy A M 2013 *Advanced Computer Science and Information Systems (ICACSIS)*, 2013 (IEEE: Bali Indonesia)
[9] Bhavsar H and Panchal M H 2012 *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1 185
[10] Min J and Lee Y 2005 *Expert Syst. Appl.* 4 603
[11] West J and Bhattacharya M 2016 *Comput. Secur.* 57 47
[12] Liu Y, Bi JW and Fan ZP 2017 *Inf. Sci.* 394 38
[13] Hsu CW and Lin SJ 2002 *IEEE Trans. Neural Netw.* 13 415
[14] Permata E, Aribowo D and Maulana A. 2014 *Semin. Nas. Teknol. Inf. dan Multimed.* 2 2.03.13
[15] Centers for Disease Control and Prevention (CDC) DPDx – Laboratory Identification of Parasitic Disease of Public Health Concern, Web <http://www.cdc.gov/dpdx/malaria/gallery.html#>