IMPLEMENTATION OF PRINCIPAL COMPONENT ANALYSIS AND SUPPORT VECTOR MACHINES METHOD TO DETECT PLASMODIUM MALARIAE IN BLOOD SAMPLE IMAGE

Kasmir I Djahari[1], Galih Hermawan[2]
[1,2] Teknik Informatika Study Program – Universitas Komputer Indonesia
Jl. Dipatiukur 112 – 114 Bandung 40132
E-mail : kdjahari@yahoo.com[1], galih.hermawan@yahoo.co.id[2]

ABSTRACT

Malaria is one of deadly diseases, especially for infants and children. Malaria is caused by the sufferer's body is exploited by the bacteria formed of Malaria Plasmodium/parasite. Malaria spreads by *Anopheles* bites, which is also known by Malaria mosquitoes. This mosquito is famous due to its role as disseminator of the Malaria parasite. Malaria Plasmodium lives in the sufferer's liver then infect blood cells. There are 2 types of plasmodium that cause malaria, it is: 1) *Falciparum Plasmodium*; 2) *Vivax Plasmodium*.

Texture analysis can be obtained by extracting the characteristic of the picture through various methods. The method using the principal component analysis and support vector machines as classification method in *Pattern Recognition*. *Pattern Recognition* is one of computer science areas, which distribute the data into particular concept it's *class* or *category*.

The conclusion of the test results that the application of the PCA and SVM method is successfully in recognition and classification of the malaria plasmodium with the accuracy of 75%.

Keywords: Pattern Recognition, Image Processing, Feature Extraction, *Principal Component Analysis*, *Support Vector Machines.*

1. INTRODUCTION

1.1 Background

Malaria is one of deadly diseases, especially for infants and children. Malaria is caused by the sufferer's body is exploited by the bacteria formed of Malaria Plasmodium/parasite. Malaria spreads by *Anopheles* bites, which is also known by Malaria mosquitoes. This mosquito is famous due to its role as disseminator of the Malaria parasite. Malaria Plasmodium lives in the sufferer's liver then infect blood cells. There are 2 types of plasmodium that cause malaria, it is: 1) *Falciparum Plasmodium*; 2) *Vivax Plasmodium*, Detection of malaria in a person's body can by microscopic observation or light microscope observation. The blood is drawn from malaria suspected patients and observed in the laboratory by the experts using light microscopy [1]. But in a way it would requires microscopic trained experts and takes a long time to get the examination result. Futhermore, most labs of Indonesia isolated areas have been inadequate so it effected to reliability of the low examination results.

With the technology and computer based system, the treatment can be performed with the *image processing technology* or digital *image processing* quickly. In the *medical* field, *image processing* technology is easier for diagnose a disease, speeding up the identification process, so that can saving the times and costs. Malaria Plasmodium detection process can be done by taking the sample of blood to each patients who has symptoms of parasites/malaria plasmodium that have been processed by using a microscope. The sample of the blood will be processing, so that the data obtained can be analyzed to detecting malaria plasmodium disease. *Image processing* technology can be obtained through by various methods to extracting the characteristics of PCA (*Principal Component Analysis*), *Watershed method*, the method LSE (*Least Square Estimation*) and Fuzzy Logic. After performing the characteristics extraction of the image, we need the classifying of the image by using several methods such as methods of SVM (*Support Vector Machines*), the method of k-NN (*k-Nearest Neighbour*). In the study [2] the introduction of blood disease using *Fuzzy Logic* method which the introduction trained for recognizing the blood diseases based on the conceiving blood image and FIS (*Fuzzy Inference System*) method to its classification with accuracy value amounted to 66.67%, and also in research [3] the blood diseases introduction using image processing technique and artificial neural texture for the classification in obtained the best classification with the average value = 65%.

Principal Component Analysis is one of the characteristic extraction methods that is capable in producing more accurate value of image characteristics by reducing the data variable input

dimensions to the main component with smaller dimension by lossing the minimum information [4]. *Support Vector Machines* Method is one of the classification methods that used in Pattern Recognition. *Pattern Recognition* is one of computer science areas, which distribute the data to particular concept of old definition. This particular concept called class or category. In study [5] is using *Support Vector Machine* in classifing the human skeleton pose of produces the accuracy till 90.67%.

In this study, by implementing *The Principal Component Analysis and Support Vector Machines* method is expected to produce more accurate of the malaria plasmodium analysis, which is recognizing the malaria plasmodium species more accurately.

1.2 Purpose and Aims

The achieved aims in this study are:

1. Identify the malaria plasmodium.
2. Determine the time of the system in malaria plasmodium detection in a blood sample.
3. Determine the system accuracy by using *Principal Component Analisys and Support Vector Machines* method in Malaria Plasmodium detection

1.3 Scope of Problems

The boundaries of the problems is:

1. Blood samples image data by light microscope.
2. The dataset was taken at PROF. DR. H. ALOE SABOE Gorontalo hospital.
3. The used file in application by extensions: *.jpg, *.jpeg.
4. The classification process using *support vector machine* method with two different plasmodium types; vivax plasmodium, falciparum plasmodium.

2. BASIC THEORIES

2.1 Computer Vision

Computer Vision is rapidly growing field devoted to analyze, modify, and high-level image understanding. The aim is to determine what's happening in front of the camera and use that understanding to control the computers, robotic systems, or giving more informative new image than the original camera image. Area application for computer vision technology including video surveillance, biometrics, automotive, photography, film production, web searches, drugs, augmented reality games, new users interface, and more [9].

2.2 The Basic Theory Of Digital Image

On image processing field, the image processing is digital image, that is continuous image that has been converted into a discrete form, both spatial coordinates and light intensity (brightness) through sampling and quantization process [10]. The image interpreted as continuous function of light intensity (x,y) on two-dimensionAL, to (x,y) expresses spatial coordinate and the value f of any points (x,y) expressed the intensity or brightness level or gray degrees (brightness/gray level).

Pixel (0,0) is located on the upper left of the image sides, the x index moves to the right and y index moves downwards. Mathematically the equation for the intensity function f (x, y) is:

$$0 \leq f(x,y) < \infty \quad (1)$$

Suppose f is a two-dimensional digital image of size MXN. So that, the representais of f in a matrix can be seen in the following figure.

$$f(x,y) = \begin{matrix} f(0,0) & f\ 0,1 & \dots & f(0,N-1) \\ f(1,0) & f(1,1) & \dots & f(1,N-1) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-0) & f(M-1) & \dots & f(M-1,N-1) \end{matrix}$$

Figure 1 Digital Image Matrix

2.3 RGB Image

RGB is color model consists of red, green, and blue that combined and formed in wide array of colors. Each primary colors has values range between 0 to 255 [11].

2.4 Principal Component Anylisis Method *(PCA)*

Principal Component Analysis is statistical technique that has been used widely both facial recognition and pattern recognition of image. Principal Component Analysis (PCA) method was first created by statisticians and was discovered by Karl Pearson on l90l which is used in the field of biology. Then there are no new developments on this technique and its development has started rapidly in the late l930 and early 1940. Principal component analysis is a statistical technique to change from most of the original variables used and correlated with each other into a set of new smaller variables and not correlated [4]. Every 4 multivariate measurement (or observation), the main component is a linear combination of the variable *p*. The main objective principal components analysis is to reduce the dimensions of the variables are interconnected and quite a lot of variable making it easier to interpret such data [4]. The method used is to determine the major components to perform over a wide orthogonal or form a linear combination Y = A 'X [6]. From here will have some key components that can provide the majority of the total diversity of the data originally. Steps - steps to perform PCA outline is as follows:

a) Standardize the data - the data imaginable. There are times when the data - the data that we get has a different scale which is quite striking. One data ranges from 10-1, while the other is worth 102. The difference of this scale will lead to asymmetries in the distribution (variance)

data. Therefore, these data should be standardized inter alia, by the following:

$$x_{baru} = \frac{x_{lama} - nilai\ rata{-}ratanya}{standar\ deviationnya} = \frac{x_i - \bar{x}}{\sigma} \quad (3)$$

b) Create a covariance matrix of variables - variables that exist Covariance is a variance between the two variables, such as X and Y. With reference to the formula of the variance, the covariance is calculated as follows:

$$variance\ (x,x) = \sum_{i-1}^{N} \frac{(x_i - \bar{x})(x_i - \bar{x})}{N-1} \quad (4)$$

$$variance\ (x,y) = \sum_{i-1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1} \quad (5)$$

c) The more variables involved, the greater the covariance matrix - its variance. Find eigen value and eigen vector of covariance matrix. Value of eigen value determines the principal component is involved. This value is calculated by solving the following equation:

$$\det[A - \lambda I] = 0 \quad (6)$$

Where:
A=Covariance Matrix
λ= eigen value
I=Identity matrix

From the equation above, the number of eigen value will be as many as the number of variables that are involved. Then, eigen vector is calculated by solving the following

$$[A - \lambda I\ IX] = 0 \quad (7)$$

X = eigen vector for eigen value is used.

The above equation solved for every eigenvalue who gained. Eigenvector with the largest eigenvalue is called the first principal component, and so on. Typically, the two principal component is enough to summarize approximately 80-90% variations in the data.

2.5 Support Vector Machines Method

Support Vector Machines (SVM) is a integrated classification method (*supervised*) because on the training process, required specific learning target. SVM is algorithm works using nonlinear mappings for changing the original training data to higher dimension. In this case, new dimension will seek the *hyperplane* to linear separated and right nonlinear mapping to higher dimension, two classes data can be separated by the *hyperplane*. SVM finds it using by *support vector hyperplane and margin* [12].

SVM appeared in 1992 by Vladimir Vapnik with Bernhard Boser and Isabelle Guyon. Basis for SVM has existed since 1960 (including early work of Vapnik and Alexei Chervonenkis on statistical learning theory).

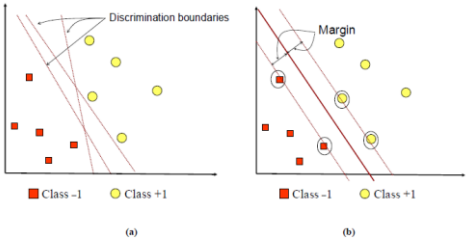


Figure 2 SVM try finding the best hyperplane which separated two classes -1 and +1

Best Hyperplane separation between two classes can be found by measuring the *margin hyperplane* and looking for the maximum points. Margin is the distance between the hyperplane to the nearest pattern of each class. The closest pattern is called *support vector*. Solid line in Figure 2.7-b shows the best hyperplane, that is located in the middle of the second class, while the red dots and yellow are in black circle as support vector. Effort to find the hyperplane is the core of the learning process on SVM [12].

SVM concept simply explained as an attempt to find the best hyperplane which serves as separator of two classes on the input space. Figure 2.3-a shows some pattern as member of two classes: +1 and -1. Pattern belonging to the class -1 symbolized by the color red (box), while the pattern on the class +1, symbolized by the yellow color (circle). Problem classification translated by trying of finds a line (hyperplane) that separates between two groups.

Various divide line alternatives (discrimination boundaries) are shown in Figure 2.3 (a). Hyperplane best separation between two classes can be found by measuring the hyperplane margin and looking for maximum points. Margin is the distance between the hyperplane to the nearest pattern of each class. The closest pattern is called a support vector. Solid line in Figure 2.3 (b) shows the best hyperplane, which is located right in the middle of the second class, while the red dots and yellow are in a black circle is a support vector. Attempts to locate hyperplane is the core of the learning process of SVM. Hyperplane defined by the following equation.

$$g:\ w^t + b = 0 \quad (8)$$

Data provided is denoted as $(x_l)^{\rightarrow} \in \Re d$ while each label is denoted yi $\in \{-1, +1\}$ for $i = 1,2,3, ..., l$ which l is the number of data. The second class is assumed to -1 and +1 separated completely by hyperplane dimension d, which is defined

$$w^{\rightarrow}.x^{\rightarrow} + b = 0 \quad (9)$$

Pattern $(x_l)^{\rightarrow}$ which includes class -1 (negative samples) can be formulated as a pattern that satisfies inequalities,

$$w^{\rightarrow}.(x_l)^{\rightarrow} + b \leq -1 \quad (10)$$

while the pattern $(x_l)^{\rightarrow}$ that including the classes +1 (positive samples).

$$w^{\rightarrow}.(x_l)^{\rightarrow} + b \leq +1 \quad (11)$$

The bigger margin founded by maximizing distance value between the hyperplane and the closest point, which is $1/\| w^{\rightarrow}\|$. It can be formulated as a *Quadratic Programming (QP) problem*, that is finding the minimum point of the equation 12 by regarding constraint equation 13.

$$\frac{min}{w} \tau(w) = \frac{1}{2}\|w\|^2 \quad (12)$$

$$y_i(\vec{x_l}.\vec{w}+b)-1 \geq 0\ ,\ \forall_i \quad (13)$$

This Problem broken bt various computation technique, there are *Lagrange Multiplier*.

$$L(\vec{w},b,a) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{l} a_i\,(yi((xi.w+b)-1))$$

i=1,2,3,….,l (14)

α_i is lagrange multiplier, valued as zero or positive ($\alpha_i \geq 0$). The optimal value of the equation 14 may be calculated by minimizing L to w and b, and maximize L against αi. By regarding to the nature that at the point of optimal gradient L = 0, the equation 14 may be modified as a maximization problem that only contains αi course, as the equation 15 below.

Maximize:

$$\sum_{i=1}^{l} ai - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j x_i x_j \quad (15)$$

Subject to:

$$a_i \geq 0(i = 1,2,3,\dots,l)\sum_{i=1}^{l} a_i y_i = 0 \quad (16)$$

From the result of this calculation is obtained αi most positive value. Data were correlated with positive αi is called support vector.

3. ANALYSIS

3.1 System Analysis

The analysis system defined as the decomposition of a whole system into its component parts to identify and evaluate the problems, barriers that occur and needs expected to be proposed the amendments.

3.2 Problems Analysis

The raising issues in this thesis is how to implement the Principal Component Analysis method as feature extraction method to be useful in recognizing and identifying the malaria plasmodium by using Support Vector Machine and also in terms of computational rapidly, producing good features characteristic and high accuracy. So, it will produce good performance application, nice and helpful.

3.3 Algoritma Analysis

This study uses a feature extraction method Principal Component Analysis and Support Vector Machines method in terms of detection. In this system the analysis phase is devoted in the process of malaria plasmodium detection, recognition chart as a whole can be seen in Figure 3.1

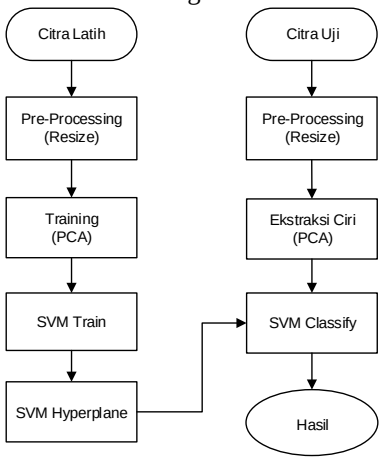


Figure 3 Alur Proses Sistem

3.4 Pre-Processing Image Analysis

Pre-Processing is the process of resizing by resizing the original RGB image to a size of 400 x 400 pixels. Resize stage is necessary for feature extraction process to equalize the pixel values of each images dataset. The process can't did well if the pixel values of each image is different.

3.5 PCA Training Analysis

PCA training is intended for reading data from the image that has been resizing of image pre-processing process to produce the vector values of the training image.



Figure 4 PCA Training Process Flow

3.6 Support Vector Machines Training Analysis

After getting a hallmark of the results of PCA train then set classes that will be used by using equations 12 and 13. After establishing the classes that will be used later to find data which will be used as support vector with equations 16 and then seek hyperplane to attract some lines to separated two pieces of the class specified by equation 11, then selected the best hyperplane with *support vector* within each class with a large margin hyperplane.

4. IMPLEMENTATION

4.1 Hardware Implementation

The hardware used to implement the system can be seen in Table 1.

Table 1 Hardware Specifications

Num.	Hardware	Spesification
1.	Processor	Intel i3 2.30 GHz
2	Monitor	Monitor 14" 1366 x 768 (32Bit) (60Hz)
3	VGA	512Mb
4	Memory (RAM)	2GB
5	Keyboard dan Mouse	

4.2 Software Implementation

The software used to build and implement the system can be seen in Table 2.

Table 2 Software Specifications

Num.	Software	Spesification
1.	Operation System	Windows 7, 8, and 8.1
2	Tools	Matlab R2009a and all
3	Programming Language	Matlab

4.3 System Examination

Testing is an important part in the development cycle of software.The purpose of testing is to find errors and flaws in software. Testing is also intended to ensure the software that built to have reliable quality as capable of reviewing the principal of the specification, analysis, design, and coding of the software itself.

4.4 Performance Examination

The results of performance testing based on the speed of time and accuracy in the process of training and testing of data with a total of 16 training dataset of the fastest and accurate according to PCA parameters, radius and sampling point. The test is as follows.

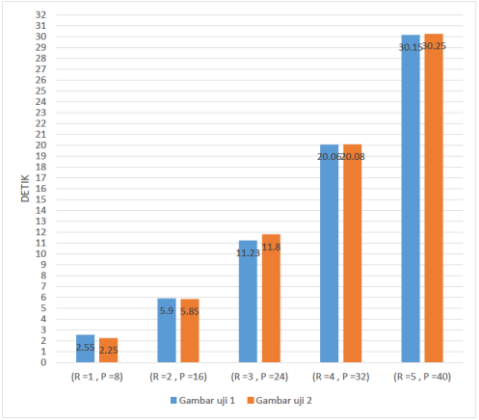


Figure 5 Graph process time fitur exraction PCA

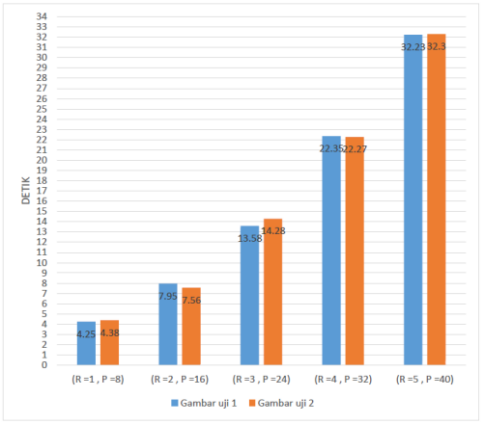


Figure 6 Graph overall time computing

Tests conducted on two test images whose image resolution of 400 x 400 pixels and 8 bits using the PCA parameter calculation process time per second.

4.5 Accurate Examination

Results of testing the accuracy of recognition knuckle arms against the following picture:

a. Testing based on sample images 1, 2, 3 & 4 malaria Plasmodium and use the best PCA parameter is r = 1 & p = 8.

Table 3 Test results based dataset

N0	**Test Images**	**SVM Clarification Result**	**SVM Testing Result**	**Description**
1	P.Falciparum Gametocyte 1.jpg	P.Falciparum	P.Vivax	Gagal
2	P.Falciparum Gametocyte 2.jpg	P.Falciparum	P.Falciparum	Berhasil
3	P.Falciparum Gametocyte 3.jpg	P.Falciparum	P.Falciparum	Berhasil
4	P.Falciparum Gametocyte 4.jpg	P.Falciparum	P.Vivax	Gagal
5	P.Falciparum Gametocyte 4.jpg	P.Falciparum	P.Falciparum	Berhasil
6	P.Falciparum Trophozoite1.jpg	P.Falciparum	P.Falciparum	Berhasil
7	P.Falciparum Trophozoite2.jpg	P.Falciparum	P.Falciparum	Berhasil
8.	P.Falciparum Trophozoite3.jpg	P.Falciparum	P.Falciparum	Berhasil
9.	P.Falciparu	P.Falciparum	P.Falciparu	Berhasil

	m Trophozoite4.jpg		m	
10.	P.Falciparum Trophozoite4.jpg	P.Falciparum	P.Falciparum	Berhasil
11.	P.Vivax Gametocyte 1.jpg	P. Vivax	P. Vivax	Berhasil
12.	P.Vivax Gametocyte 2.jpg	P. Vivax	P.Falciparum	Gagal
13.	P.Vivax Gametocyte 3.jpg	P. Vivax	P. Vivax	Berhasil
14.	P.Vivax Gametocyte 4.jpg	P. Vivax	P. Vivax	Berhasil
15.	P.Vivax Gametocyte 4.jpg	P. Vivax	P. Vivax	Berhasil
16.	P.Vivax Trophozoite1.jpg	P. Vivax	P.Falciparum	Gagal
17.	P.Vivax Trophozoite2.jpg	P. Vivax	P. Vivax	Berhasil
18.	P.Vivax Trophozoite3.jpg	P. Vivax	P.Falciparum	Gagal
19.	P.Vivax Trophozoite4.jpg	P. Vivax	P. Vivax	Berhasil
20.	P.Vivax Trophozoite4.jpg	P. Vivax	P. Vivax	Berhasil

b. Testing based on input data that one of or not suitable with the trained data.

Table 4 Testing Result based on wrong data

Num.	Testing Images	Result	Description
1.	Mangga.jpg	Failed	Because the picture pixel unsuitable and different objects
2.	Durian.jpg	Failed	Because the picture pixel unsuitable and different objects
3.	Manggis.jpg	Failed	Because the picture pixel unsuitable and different objects
4.	Melon.jpg	Failed	Because the picture pixel unsuitable and different objects
5.	Anggur.jpg	Failed	Because the picture pixel unsuitable and different objects

Based on testing accuracy result of which has been employed with the following data:

Many Data Training : 20
Many Data Testing : 20
Many True Data : 15
Many Data One : 5
Percentage =Lots of Data True/Many Data Testing * 100%
= 15/20 * 100%
= 75%

So from these data we can take the conclusion that the percentage of success is done by the system is 75%.

5. THE END

5.1 Conclusion

Based on the discussion of the analysis, design, implementation to the stage of testing, it can be concluded as well as suggestions for the Principal Component Analisys Methods Implementation and Support Vector Machines in detecting malaria plasmodium in the blood sample image.

Based on a comparison between the objectives of software development with the results of the implementation and testing, it could be concluded as follows:

1. The application has been successfully implemented in the detection of malaria plasmodium in the blood sample image, making it easy to identify and detect these types of malaria only image blood samples have been taken using a light microscope.
2. From the test results, the best system performance parameters contained in the PCA with radius = 1 and the sampling point = 8 producing process faster computation time and also more accurate than PCA parameters.
3. The results of testing the accuracy of the implementation of the method of Principal Component Analysis and Support Vector Machines in the recognition and classification of malaria plasmodium produces an accuracy of 75%.

5.2 Recommendation

Apart from the advantages of a software must have flaws, and therefore for developing a more suggestions can be given as follows:

1. Develop an application recognition and classification of malaria plasmodium in real time.
2. Increase the number of datasets for training around 50.
3. Using the hardware support is more reliable in order to get maximum results in using Principal Component Analysis Idan Support Vector Machines Method.
4. Application of this method can be developed in a different case not only in the field but other malaria Plasmodium.
5. For further research may eventually feature extraction method that is more satisfactory.

BIBLIOGRAPHY

[1] A. Arsin, ”Analisis Perilaku Masyarakat Terhadap Kejadian Malaria di Pulau Kapoposang kabupaten Pangkajene Kepulauan,” Jurnal Kedokteran dan Farmasi MEDIKA, Jakarta, 2003.

[2] D. L. Lestari, “Pengenalan Penyakit Darah Dengan Citra Darah Menggunakan Metode Logika Fuzzy,” Jurnal Ilmu Komputer dan Informasi Universitas Indonesia, Depok, 2008.

[3] A. R. Praida, “Pengenalan Penyakit Darah Menggunakan Teknik Pengolahan Citra Dan Jaringan Syaraf Tiruan” Jurnal Ilmu Komputer Universitas Indonesia, Depok, 2008

[4] R. Firdaus, “Pengenalan Pola Tanda Tangan Dengan Menggunakan Metode Principal Component Analysis (PCA),” Jurnal Ilmu Komputer Universitas Dian Nuswantoro, Semarang, 2013.

[5] S. N. Endah dan M. R. Widyanto, “Klasifikasi Pose Skeleton Pada Manusia dengan Support Vector Machines,” Jurnal Ilmu Komputer dan Informasi Universitas Diponegoro, Semarang, 2012.

[6] R. S. Pressman, Ph.D, Software Engineering A Practitioner's Approach, 6th penyunt., New York: McGraw - Hill Higher Education, 2005.

[7] Nurhari, Ogi ”*Plasmodium Sp*,” 2009. [Online] Available: scribd.com/doc/51574461/Epidemologi-Malaria [Diakses 20 Aktober 2015].

[8] Drs. Slamet Prawirohartono, Prof. Dr. Suhargono Hadisumarto, “Sains Biologi-2a” Jakarta, Bumi Askara, 2000.

[9] Fadlisyah, “Computer Vision dan Pengolahan Citra,” Andi, Yogyakarta, 2009.

[10] R. Gonzalez dan R. Wood, Digital Image Processing, New Jersey: Prentice Hall International, 2008.

[11] D. Putra, “Pengolahan Citra,” Andi, Yogyakarta, 2010.

[12] P. E. Yunanto, T. Agung dan E. Rachmawati, “Pengenalan Telapak Tangan Melalui Citra Digital Dengan Menggunakan Metode Eigenpalms dan SVM,” Jurnal Science Test TELKOM, 2012.

[13] P. P. Widodo, R. T. Handayanto dan Herlawati, “Penerapan Data Mining Dengan Matlab,” Bandung: Rekayasa Sains, 2013.

[14] A. Rosa dan M. Shalabudin, “Modul Pembelajaran Rekayasa Perangkat Lunak (Terstruktur dan Berorientasi Objek),” Modula, Bandung, 2011.

[15]G. A. Away, The Shortcut of MATLAB Programing, Bandung: Informatika Bandung, 2006.