









ORIGINAL

Classification of Malaria Parasite Plasmodium Falciparum Based on Blood Smear Images Using Support Vector Machine Approach

Clasificación del parásito de la malaria Plasmodium falciparum basada en imágenes de intercambio de sangre utilizando un enfoque de máquina de vectores de soporte

Nur Chamidah^{1,2}  , Toha Saifudin^{1,2} , Riries Rulaningtyas³ , Adam Anargya Mawardi³ , Puspa Wardhani⁴ , I Nyoman Budiantara⁵ , Naufal Ramadhan Al Akhwal Siregar¹ 

¹Airlangga University, Department of Mathematics, Faculty of Science and Technology. Surabaya 60115, Indonesia.

²Airlangga University, Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology. Surabaya 60115, Indonesia.

³Airlangga University, Department of Physics, Faculty of Science and Technology. Surabaya 60115, Indonesia.

⁴Airlangga University, Department of Clinical Pathology, Faculty of Medicine. Surabaya 60115, Indonesia.

⁵Sepuluh Nopember Institute of Technology, Department of Statistics, Faculty of Sciences and Data Analytics. Surabaya 60111, Indonesia.

Cite as: Chamidah N, Saifudin T, Rulaningtyas R, Mawardi AA, Wardhani P, Budiantara IN, et al. Classification of Malaria Parasite Plasmodium Falciparum Based on Blood Smear Images Using Support Vector Machine Approach. Data and Metadata. 2025; 4:568. <https://doi.org/10.56294/dm2025568>

Submitted: 08-05-2024

Revised: 01-09-2024

Accepted: 06-12-2024

Published: 01-01-2025

Editor: Adrián Alejandro Vitón Castillo 

Corresponding Author: Nur Chamidah 

ABSTRACT

Malaria remains a significant global health problem, especially in tropical and subtropical regions. The disease results in a large number of clinical cases and deaths each year, with high-risk groups including infants, toddlers, and pregnant women. Accurate and rapid diagnosis is a key factor in treating this disease. To address this problem, this research aims to develop an automatic system for classifying the malaria parasite Plasmodium Falciparum based on blood smear images. The method used includes image feature selection using Principal Component Analysis (PCA) and the Support Vector Machine (SVM) approach for classification. The results showed that in the image feature selection process, the normal malaria category showed typical characteristics with PC1 and PC2 values that tended to be negative and scattered, while the parasitic malaria category showed greater variability in the PC1 and PC2 components. Furthermore, evaluation of the accuracy of the classification system using SVM with three different kernel types shows promising results. The average accuracy through K-fold cross-validation for the polyinomial, linear, and radial basis function kernels is 96,7 %, 98,9 %, and 94,4 %, respectively. These results highlight the significant potential of utilizing SVM in the classification of the malaria parasite Plasmodium Falciparum based on blood smear images.

Keywords: Malaria; Parasite Classification; Principal Component Analysis; Support Vector Machine.

RESUMEN

La malaria sigue siendo un importante problema sanitario mundial, sobre todo en las regiones tropicales y subtropicales. La enfermedad provocaba un número considerable de casos clínicos y muertes cada año, y entre los grupos de alto riesgo se encontraban los lactantes, los niños pequeños y las mujeres embarazadas. El diagnóstico preciso y rápido era un factor clave en la gestión de la enfermedad. Para abordar este problema, la investigación se propuso desarrollar un sistema automatizado de clasificación de parásitos de malaria Plasmodium falciparum basado en imágenes de frotis sanguíneos. Los métodos empleados incluyeron la selección de características de la imagen mediante el análisis de componentes principales (ACP) y el enfoque de la máquina de vectores de apoyo (MVA) para la clasificación. Los resultados de la investigación indicaron que, en el proceso de selección de características de la imagen, la categoría de malaria normal

presentaba características distintivas con valores de PC1 y PC2 que tendían a ser negativos y dispersos, mientras que la categoría de malaria parasitaria mostraba una mayor variabilidad en los componentes PC1 y PC2. Además, la evaluación de la precisión del sistema de clasificación mediante MVA con tres tipos de kernel diferentes mostró resultados prometedores. La precisión media mediante validación cruzada K-fold para los núcleos polinomial, lineal y de función de base radial fue del 96,7 %, 98,9 % y 94,4 %, respectivamente. Estos resultados pusieron de relieve el importante potencial de la utilización de MVA en la clasificación de parásitos de la malaria *Plasmodium Falciparum* basada en imágenes de frotis sanguíneo.

Palabras clave: Malaria; Clasificación de Parásitos; Análisis de Componentes Principales; Máquina de Vectores de Apoyo.

INTRODUCTION

Malaria is one of the most common diseases in tropical and sub-tropical areas. This disease can cause death, especially in groups of people at high risk, namely babies, toddlers and pregnant women. Malaria is caused by the *Plasmodium* parasite resulting from the bite of the female *Anopheles* mosquito.⁽¹⁾ Malaria is still one of the main health problems in the world and it is reported that 3,2 billion people in the world are infected with malaria.⁽²⁾ The risk areas for malaria transmission come from 108 countries, and it is estimated that there are around 300-500 million clinical cases of malaria worldwide with a mortality rate of more than 1 million people per year.⁽³⁾

In Indonesia, it is estimated that around 46,2 % of the 210,6 million total population live in malaria endemic areas and 56,3 million people live in moderate to high risk areas. More than 3 million clinical cases of malaria are reported each year, mainly in poor areas, and 30 000 cases of malaria deaths are reported by health service units, including community health centers and hospitals.⁽⁴⁾ In the process of treating malaria, one of the main stages is the patient's diagnosis. A good standard for diagnosing malaria is to carry out a microscopic examination of thick smear or thin smear images of the patient's blood samples, to be examined by competent medical personnel.⁽⁵⁾ If *plasmodium* parasites are detected in the blood, then the sample is classified as infected, and the patient is classified as suffering from malaria and must undergo further treatment.

The main problem with this diagnostic method is that microscopic examination of samples often cannot be carried out directly, due to the limited availability of qualified medical personnel to carry out visual examinations of thick smear or thin smear images of blood samples. Therefore, a solution is needed to overcome this problem, one of which is to check the presence or absence of *plasmodium* parasites in blood samples automatically, with the help of technology. In particular, image classification technology can be used for automated examination of blood samples.^(5,6,7,8) Thus, accurate early detection of malaria is very important.

Commonly used malaria detection techniques include calculating the percentage of normal red blood cells (erythrocytes) and the percentage of blood cells infected with malaria.⁽⁹⁾ Manually counting blood images obtained from a light microscope is a technique that is often used, but this technique requires a long and laborious process and requires experts to carry out the calculations.⁽¹⁰⁾ The World Health Organization (WHO) recommends the use of method called parasite-based diagnostic testing, for example microscopic analysis. However, this method requires an experienced and competent microscopist. Apart from that, the large amount of data that must be analyzed in a short time by microscopists will also be a problem in Mass Blood Surveys (MBS). Another problem is the unavailability of medical equipment in rural areas. These problems can affect the determination of the type of parasite, resulting in delays in patient treatment.⁽¹¹⁾

Of the various existing problems, Computer Aided Diagnosis (CAD) can be used to detect malaria early. CAD is a system used to help interpret medical images in a short time and improve the accuracy of diagnosis. Several computer-aided parasite identification studies have been conducted using various experimental methods. Several methods used to segment images include otsu, threshold, k-means, active contour, adaptive color, and edge detection.^(12,13,14,15,16,17) Then, to classify parasites, there are several methods that can be used, including Support Vector Machine (SVM), K-Nearest Neighbor, Naïve Bayes, Decision Tree, Fuzzy, Artificial Neural Network, and Penalized Spline Nonparametric Poisson Regression.^(18,19,20,21,22)

A research on *plasmodium* stage identification was conducted by Muhimmah et al.⁽²³⁾ who identified *Plasmodium Ovale* using the Support Vector Machine (SVM) method using MATLAB software, where the manual cropping was carried out for 31 images with a size of 200 x 200 pixels as a pre-processing stage. Next, the data was segmented using the threshold method to identify the *Plasmodium Ovale* parasite. Then proceed with extraction and selection of shape, size and texture features to classify the image based on its stages. Of the 12 attributes resulting from feature selection used for 12 images in the classification system testing stage using the multi-class SVM method, the accuracy value was 83,3333 %.⁽²³⁾ Furthermore, a research conducted by Ramadhan et al.⁽²⁴⁾ which classified Malaria data using the Support Vector Machine (SVM) method where detection and

classification of severe malaria is carried out based on the patient's data examination history using the SVM method with previously using a normalization technique using min-max on the dataset and cross validation techniques with several K value experiments on the results. The study also compared the SVM method with Naïve Bayes (NB) where the accuracy of the SVM model is superior to Naïve Bayes with an average accuracy gap of 25 %. The accuracy produced by applying the SVM method was 92,3 %.⁽²⁴⁾

Based on the facts that the previous researches still remained several shortcomings, we therefore propose a method to detect the presence of malaria in red blood cells through the characteristics of red blood cells that suffer from malaria and red blood cells that do not suffer from malaria based on the extraction of color and texture characteristics. Extraction of color and texture features is used because infected and uninfected malaria images are most easily differentiated through color and texture characteristics in the image, so these features can provide important information about the characteristics and patterns contained in malaria images and are expected to produce good accuracy in classification. Classification using Support Vector Machine (SVM) is used because it has several advantages and is in accordance with the characteristics of malaria image data. Classification is carried out based on the characteristics of each stage class to assist laboratory and medical experts in carrying out early detection of malaria which is also expected to support further research.

METHOD

In this section we provide research methods which include a series of steps starting from data collection, research variables, and research steps.

Data Collection

In this research, we use primary data taken through a practicum process carried out at the Physics Laboratory, Faculty of Science and Technology, Airlangga University, Indonesia. The data used in this research are images of the malaria parasites *Plasmodium Falciparum* obtained from the Department of Clinical Pathology, Airlangga University.

Research Variables

The independent variable used in this research is the image of plasmodium red blood cells, both those detected by malaria and those not detected by malaria, obtained from direct microscopy of blood samples from malaria patients. The dependent variable used in this research is the accuracy of malaria image classification results using the SVM method with accuracy metrics on separate test data. The control variable used in this research is a microscope magnification of 1000x at the time of collection which is controlled by calibrating and maintaining consistent magnification. The amount of training data and test images is controlled by dividing the data randomly and proportionally.

Research Steps

The research steps starting from the literature review to the data analysis process are presented in figure 1. Next, steps of data analysis and steps of digital image processing are given in figure 2 and figure 3, respectively.

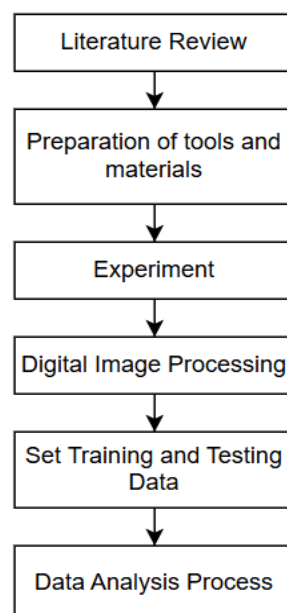


Figure 1. Steps of Research

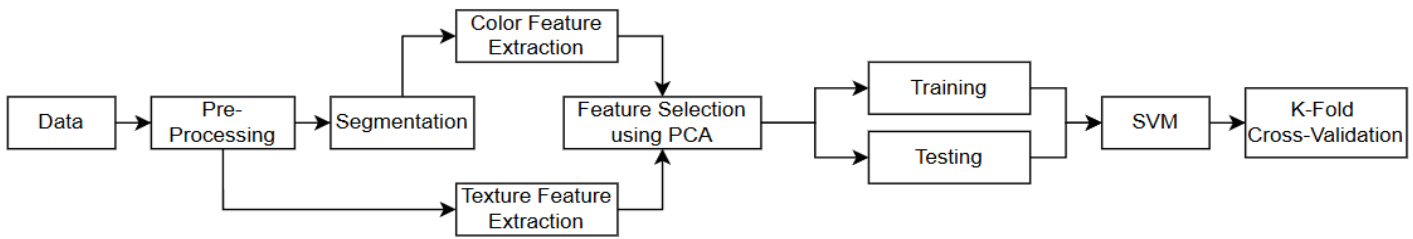


Figure 2. Steps of Data Analysis

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a learning system that hypothesizes using linear functions in high-dimensional space and is trained with algorithms based on optimization theory by applying learning biases derived from statistical theory.⁽²⁵⁾ The SVM is a supervised machine learning that can solve various problems such as text categorization, handwriting, digit recognition, tone recognition, image classification and object detection, as well as data classification.⁽²⁶⁾

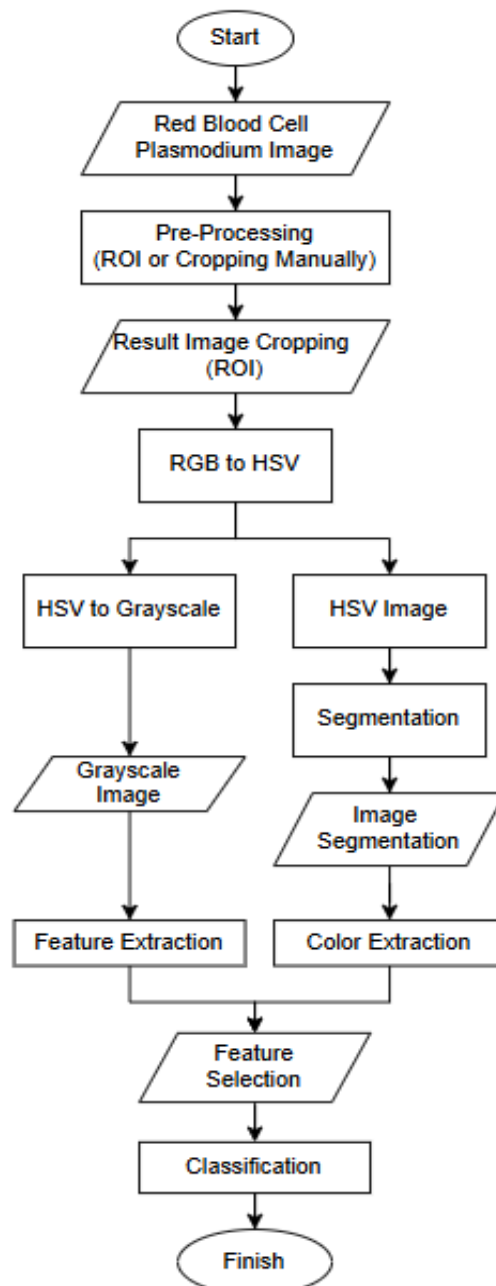


Figure 3. Steps of Digital Image Processing

The data on the boundary plane are called support vectors. Two classes can be separated by a pair of parallel boundary planes. The first limiting field limits the first class while the second limiting field limits the second class. Classification models look for decision boundaries to separate data into one class and another. If it is two-dimensional, then the decision boundary obtained is a line. In three-dimensional data, the decision boundary is a plane. In general, the concept of a dividing plane is called a hyper plane. For F-dimensional data, our decision plane has dimension F-1.

$$x_1w_1 + x_2w_2 + \dots + x_Fw_F \quad (1)$$

$$\mathbf{xw} + b = 0 \quad (2)$$

More mathematically, as in equations (1) and (2), this concept is similar to binary classification using sign and thresholding functions. If it is possible to perfectly separate two classes of data with a hyperplane (linearly separable), the choice of hyperplane that can be made is not one. This means that the boundary line can be shifted, while still separating the data perfectly.⁽²⁷⁾

Metrics Evaluation

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives).

Table 1. Confusion Matrix		
Data Class	Classified as pos	Classified as neg
Pos	True Positive (TP)	False Negative (FN)
Neg	False Positive (FP)	True Negative (TN)
Source: Sokolova et al. ⁽²⁸⁾		

These four counts constitute a confusion matrix shown in table 1 for the case of the binary classification. The evaluation metrics, namely, accuracy, precision, recall, and f1-score are given in table 2.

Table 2. Metrics Evaluation Using the Notation of table 1		
Measure	Formula	Evaluation Focus
Accuracy	$(TP+TN)/(TP+FN+FP+TN)$	Overall effectiveness of a classifier
Precision	$TP/(TP+FP)$	Class agreement of the data labels with positive labels
Recall	$TP/(TP+FN)$	Effectiveness of a classifier to identify positive labels
F1-score	$2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$	Relations Between data's positive labels and those given by a classifier
Source: Sokolova et al. ⁽²⁸⁾		

RESULTS

Digital image processing for image classification of malaria parasites *Plasmodium Falciparum* based on blood smear image with SVM approach is a series of steps that involve extracting significant information from the microscopic image of red blood cells containing the parasite plasmodium that causes malaria. This process aims to improve the identification accuracy of normal and detected malaria parasites.

At this stage, image pre-processing is carried out to process the original image by getting the desired object that can be seen more clearly and can facilitate the next process. At this stage, the original image measuring 680 x 512 pixels is carried out Region of Interest (ROI) by cropping manually to 100 x 100 pixels. The cropping process is done to mark and clarify one of the normal and malaria detected plasmodium. The position of the desired object is not required to be exactly at a certain angle, but the object must remain intact and not cut off.

The results of the cropping process for normal malaria image objects and parasitic malaria image objects as shown in figures 4 (c) and 4(d) will be used for the cropping process. This image will be used for the next pre-processing stage, namely the image will be converted to color space, where the image will be converted to the HSV color space to increase the image contrast and obtain the base color of the image.

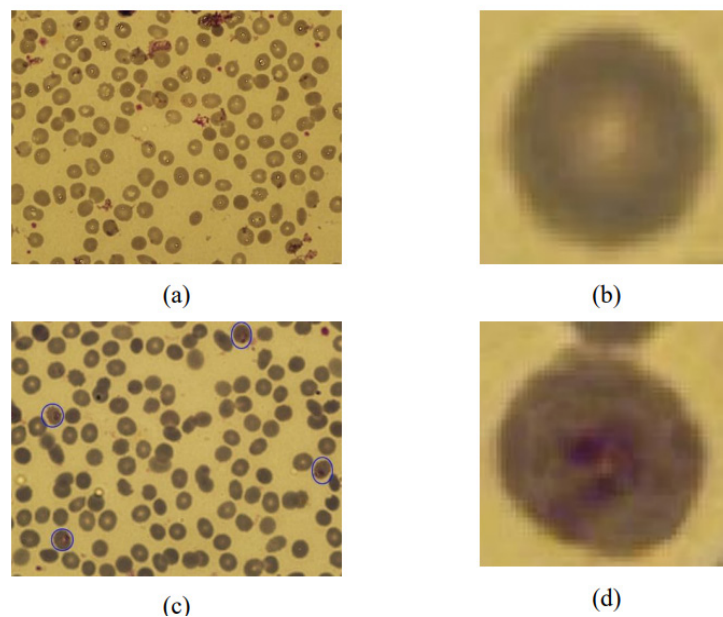


Figure 4. Cropping Image Pre-Processing, (a) Normal Before Cropping, (b) Normal After Cropping, (c) Parasite Before Cropping, and (d) Parasite After Cropping



Figure 5. Converting Image RGB to HSV, (a) RGB, (b) HSV

The results of the color space conversion process from RGB to HSV shown in figure 5 (a) to (b) will then be further processed to the next stage. The HSV image will be segmented first to separate the object with the background for the color feature extraction stage and the HSV image will be converted to grayscale color space for the texture feature extraction stage. The results of converting the HSV color space to grayscale are then subjected to contrast normalization to get more contrast in the image which can facilitate and get maximum results in the next stage, namely the extraction of texture features.



Figure 6. Converting Image Color HSV to Grayscale

The next stage is image segmentation, which will be carried out for the analysis of normal malaria images and parasitic malaria images in color feature extraction only. This segmentation process is done to separate the object from its background so that the color in the background of the image is not extracted at the color feature extraction stage, so that it will get more accurate results in the next analysis. In this segmentation stage, a morphology operation is also used to remove noise from the image with the opening operation method used to remove small dots in the white area as shown in figure 7.



Figure 7. Image Segmentation, (a) Before segmentation (b) After Segmentation

The results of image processing after going through this segmentation stage will be input or input to the color feature extraction process. Where the calculation will focused on the color characteristics of the image through several parameters. parameters. The segmentation process in this study uses the K-means clustering method by dividing the image into several parts (clusters) based on color similarity, the number of clusters used in this study is two clusters ($k = 2$).

After a digital image processing processes at the preprocessing and segmentation stages, the feature extraction process will be carried out. The feature extraction used in this research is color feature extraction and texture feature extraction. This color feature extraction aims to determine the results of the analysis on the RGB (Red, Green, Blue) color channel in the image through several test parameters. Based on the results of color feature extraction with a total of 50 normal malaria image data and 50 parasitic malaria images, the results obtained will be averaged for each parameter shown in table 3 below.

Table 3. Results of Color Feature Extraction					
Parameter	Colors	Result		Range	
		Normal	Parasite	Normal	Parasite
Mean	Red	19,7252	20,9012	18,6518 - 20,5564	17,4875 - 25,1139
	Green	122,9710	124,8302	112,7572 - 132,3880	106,1836 - 145,1676
	Blue	152,4209	152,9191	138,8878 - 166,2085	133,4296 - 171,0103
Standard Deviation	Red	2,0705	19,9871	1,5710 - 3,4129	3,1356 - 37,5689
	Green	10,9928	12,8580	7,6451 - 18,8808	6,6728 - 21,9265
	Blue	37,7055	41,6300	30,3088 - 46,4773	33,7293 - 48,9954

From the description on table 3, it can be concluded that the main characteristics of the color features of normal malaria images and parasitic malaria images are not much different in the Mean value, it's just that there is a slight difference in Standard Deviation where parasitic malaria images have a slightly higher value than normal malaria images in the Green and Blue color with an average difference of 1,8652 and 3,9245, while there is a contrasting difference in Standard Deviation in the Red color channel with an average difference of 17,9166.

Furthermore, it will be carried out to the extraction of texture features. It aims to obtain information about the texture pattern in the image with several test parameters. In this study, the extraction of texture features used is the GLCM (Gray Level Co-Occurrence Matrix) method which is used to characterize the texture of the image and count how often pairs of pixels with certain values and certain spatial relationships in the image make GLCM, and then extract the statistical size of the matrix. The input used is the pre-processed image of RGB to HSV color channel conversion which is then converted to Grayscale and normalized. At this stage, 4 types of test parameters will be used, namely contrast, correlation, energy, and homogeneity with the results in table 4.

Table 4. Results of Texture Feature Extraction GLCM				
Parameter	Results		Range Extraction	
	Normal	Parasite	Normal	Parasite
Contrast	200,4754	447,5323	128,8702 - 402,4850	227,1738 - 751,7972
Correlation	0,9819	0,9602	0,9646 - 0,9885	0,9312 - 0,9797
Energy	0,1335	0,1158	0,0988 - 0,1588	0,0518 - 0,1588
Homogeneity	0,6894	0,5757	0,3781 - 0,7647	0,2811 - 0,7107

Based on table 4, the results of each parameter for each image prove that the texture characteristics of normal malaria images tend to have a smoother texture, a slightly stronger gray degree, more uniform, and more homogeneous. While parasitic malaria images have a coarser texture characteristic, a slightly less strong degree of gray, less uniform, and less homogeneous.

The feature selection used in this research is to use the Principal Component Analysis (PCA) algorithm method to reduce the dimensions of a dataset. PCA is a technique used to reduce the dimensionality of a dataset while maintaining important information. The purpose of PCA is to transform data into a lower dimensional space, in this research is 2 dimensions in such a way that the new components are linear combinations of the original components in the data. The PCA results that form a new dataset with lower dimensions can be seen in table 5 below.

Table 5. Results of Feature Selection using PCA							
Index	PC1	PC2	Output	Index	PC1	PC2	Output
0	-155,7119	-6,5901	0	51	53,3512	-17,1021	1
1	-161,9929	-3,3428	0	52	139,2544	-9,6305	1
2	-185,4459	4,6149	0	53	238,4737	1,8119	1
3	-151,1463	-2,9491	0	54	26,3106	0,4461	1
4	-136,1424	-11,1488	0	55	10,1727	1,1151	1
5	-112,6838	-7,8650	0	56	-10,9888	5,7403	1
6	-99,6702	-11,8412	0	57	75,8677	-3,1441	1
7	-168,5495	11,5468	0	58	-94,7602	0,2889	1
8	-160,3219	1,6830	0	59	-6,7935	2,7270	1
9	-181,5918	3,9786	0	60	-42,6098	7,0815	1
10	-157,8474	7,4618	0	61	224,7057	-27,3378	1
11	-149,9155	-2,9692	0	62	182,7962	-1,4632	1
12	-134,7113	-9,1304	0	63	158,8862	-22,0235	1
13	-84,2897	-14,5915	0	64	108,7764	4,2907	1
14	-150,9371	1,2929	0	65	80,13064	-6,3466	1
15	-169,7988	-5,4408	0	66	357,2843	8,5841	1
16	-132,9585	1,5566	0	67	149,4918	1,8881	1
17	-145,9703	-6,8796	0	68	69,0740	-4,7007	1
18	-107,0186	-7,3451	0	69	110,9493	-6,8294	1
19	-118,9709	-6,9693	0	70	41,4494	-4,8302	1
20	-121,8407	-9,1514	0	71	147,2304	-3,8866	1
21	-170,6575	9,9528	0	72	224,0514	-2,5317	1
22	-172,2308	10,8050	0	73	94,8124	14,7624	1
23	-106,8828	-7,0098	0	74	-45,6477	6,8417	1
24	-164,5044	11,8906	0	75	-96,4008	30,7042	1
25	-133,5954	13,1984	0	76	287,5313	4,9783	1
26	-195,3708	18,1558	0	77	198,5518	7,9665	1
27	-107,9570	-5,7894	0	78	428,4936	1,2166	1
28	-143,2649	4,3022	0	79	253,5357	23,2884	1
29	14,6091	-2,4808	0	80	130,0422	28,7928	1
30	35,8638	-3,6584	0	81	108,3954	4,7465	1
31	-116,8668	1,3477	0	82	106,2412	0,5927	1
32	-45,7755	-6,5456	0	83	336,8664	8,0911	1
33	-50,1848	-4,0106	0	84	314,0506	-4,3034	1
34	-117,0158	-5,7165	0	85	182,0360	6,8588	1
35	5,7655	-5,6856	0	86	127,0103	10,2941	1
36	-145,4638	3,2684	0	87	259,9983	5,6647	1

37	77,8911	-15,0720	0	88	-14,7942	-0,4298	1
38	-77,3142	-4,6002	0	89	33,9998	19,2008	1
39	-64,9700	1,0632	0	90	122,0228	13,6472	1
40	-51,6154	2,5842	0	91	14,8341	-7,0122	1
41	-85,3115	8,4295	0	92	84,6756	-13,8516	1
42	-143,5918	-9,3643	0	93	56,9946	0,8306	1
43	-184,5405	1,8419	0	94	176,7854	3,3029	1
44	-161,1158	-8,5272	0	95	57,2766	-0,5522	1
45	-193,2802	-10,3760	0	96	40,1794	4,0024	1
46	-191,8250	5,4548	0	97	258,2667	-5,2623	1
47	-192,4267	7,9944	0	98	50,4085	0,9046	1
48	-158,4419	-5,4259	0	99	357,9537	-14,2871	1
49	-165,0753	-0,0323	0	Normal (Mean)	-123,8533	-1,3617	0
50	25,4414	-7,0512	1	Paracite (Mean)	123,8533	1,3617	1

Based on table 5, there is a pattern or structure that distinguishes the normal malaria category with output 0 and parasitic malaria with output 1 in a two-dimensional representation. The pattern or structure that distinguishes normal and parasitic malaria can be seen in the two-dimensional distribution plots in figure 8 as follows.

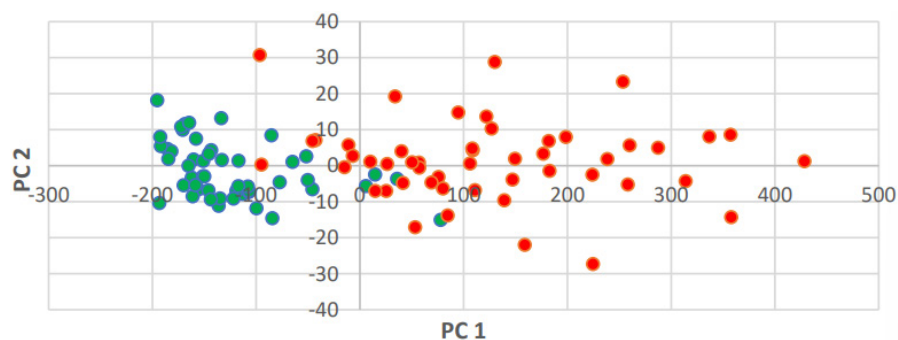


Figure 8. Distribution Plots of PCA, Normal (Green) and Paracite (Red)

Table 6. Performance Evaluation of SVM Training Model					
Kernel	C	Accuracy	Precision	Recall	F1-Score
Polynomial (Poly)	100	97,77	100	95,35	97,62
	10	92,22	90,91	93,02	91,95
	1	91,11	90,70	90,70	90,70
	0,1	90	94,74	83,72	88,89
Linear	100	100	100	100	100
	10	100	100	100	100
	1	100	100	100	100
	0,1	98,89	100	97,67	98,82
Radial Basis Function (RBF)	100	95,55	95,35	95,35	95,35
	10	93,33	91,11	95,35	93,18
	1	92,22	90,91	93,02	91,95
	0,1	90	92,50	86,04	89,16

The PCA results are used as input for the SVM classification algorithm, to build the classification model. PC1 and PC2 can be used as features to train the SVM model, which is expected to separate normal and parasitic malaria categories. In this study, the dataset will be divided into two categories, namely training set and test set, where in this study the dataset will be divided into 90 % training data and 10 % testing data. After dividing

the dataset into training and testing, hyperparameter tuning and training are then carried out, where in this study several variations of kernels will be tested, namely polynomial, linear, and radial basis function (RBF), as well as variations in C values of 100, 10, 1, and 0,1 which will be set before performing SVM classification to see which kernel and C hyperparameter tuning produces the best classification results for malaria image classification in this study. The results of evaluating the performance of the SVM training model with variations in hyperparameter tuning Cost (C) and Kernel function can be seen in table 6.

Based on the model performance evaluation results on table 6, The best hyperparameter should be a balance between accuracy and overall model performance. In this case, for the linear kernel, although the accuracy, precision, recall, and f1-score values at C=100, 10, 1 are higher, at C=0,1 are very close or even as high. Therefore, the value of C=0,1 is considered the most optimal choice because it provides a good balance between accuracy and model performance. The linear kernel with the most optimized hyperparameter C performs best with the highest accuracy, precision, recall, and f1-score results when compared to the poly kernel and RBF kernel. The poly kernel and RBF kernel are also effective by providing good results with high accuracy and balance between precision and recall, but slightly inferior to the linear kernel in terms of accuracy. The results of the comparison of the performance evaluation of the training model for each kernel with the most optimal hyperparameter C can be seen in the comparison graph of the accuracy, precision, recall, and f1-score results of each kernel in figure 9 below.

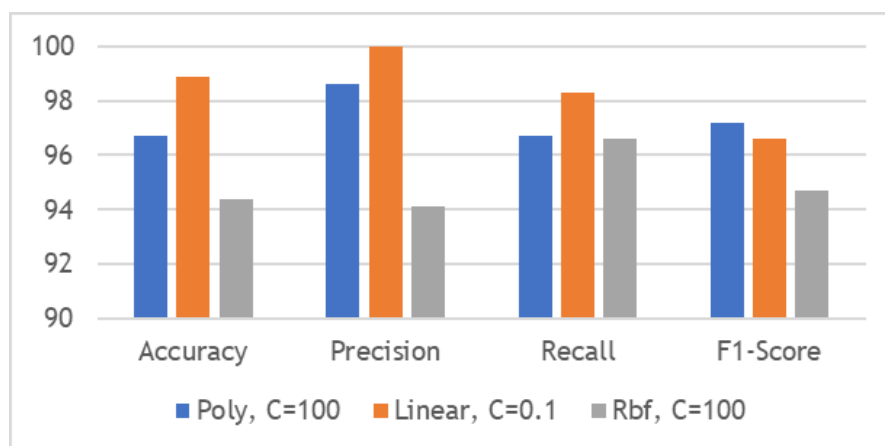


Figure 9. Comparison of Evaluation Training Model

After evaluating the training model is testing with testing data. Testing or testing Support Vector Machine (SVM) is done after getting the results of each kernel with the best hyperparameters that have been done in the previous SVM training stage. In the testing stage, the training data will be tested with testing data that has never been seen before. Each kernel with the best hyperparameter is as follows, poly kernel with hyperparameter C = 100, linear kernel with hyperparameter C = 0,1, and RBF kernel with hyperparameter C = 100, where these results are considered the best and optimal after the training process and model performance evaluation. Furthermore, calculations are carried out to print several model performance evaluation matrices such as accuracy, precision, recall, and f1-score based on the results of the confusion matrix that has been calculated in the previous stage.

The results of the model performance evaluation, namely accuracy, precision, recall, and f1-score are obtained from the calculation of the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values obtained from the confusion matrix, which are then displayed to provide a complete overview of the classification model performance of each kernel and the most optimal hyperparameters that have been trained in the previous stage. Then the visualization of the confusion matrix is shown, which is an important performance evaluation tool for classification models. The results of the SVM testing model performance evaluation matrix and confusion matrix results can be seen in table 7 and figure 10 below.

Kernel	C	Accuracy	Precision	Recall	F1-Score
Polynomial (Poly)	100	100	100	100	100
Linear	0,1	100	100	100	100
Radial basis function (Rbf)	100	100	100	100	100

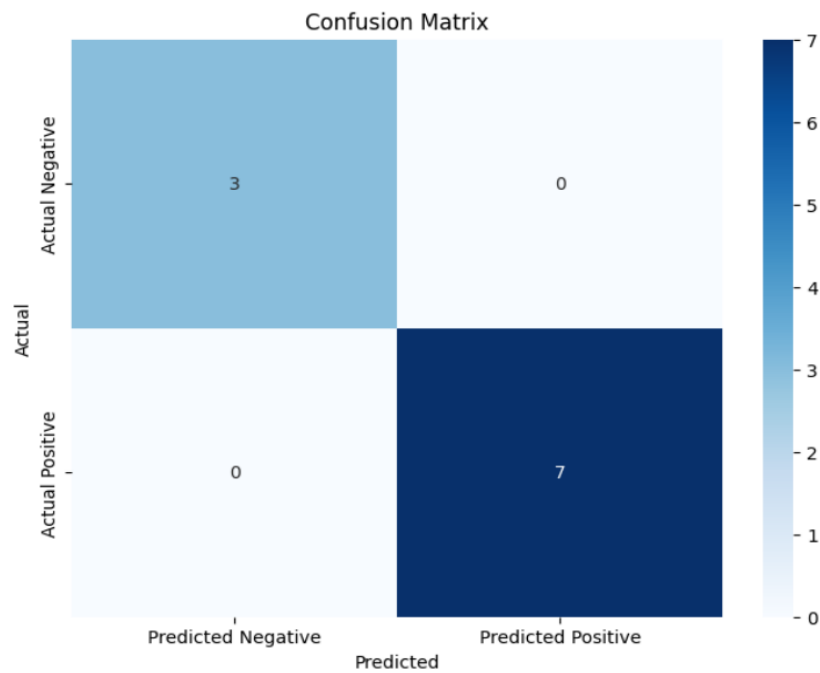


Figure 10. Confusion Matrix of Testing Model

Based on table 7 and figure 10, the accuracy, precision, recall, and f1-score values are 100 % for all model performance evaluation values on SVM classification for the testing data used. After that, from the results of the performance of training data and testing data, a validation technique is needed for the entire data with K-fold cross validation. In this research, K = 10 or the process is repeated 10 times where each fold is used as validation data alternately. K-fold cross validation in this study was carried out with each test being K-fold cross validation with all data, where the model was trained and evaluated k times. The results of K-fold cross validation with all data for poly kernel with hyperparameter C=100, linear kernel with hyperparameter C=0,1 and RBF kernel with hyperparameter C=100, can be seen in the following (see table 8).

Kernel	C	Accuracy	Precision	Recall	F1-Score
Polynomial (Poly)	100	96,7 %	98,6 %	96,7 %	97,2 %
Linear	0,1	98,9 %	100 %	98,3 %	99,1 %
Radial basis function (RBF)	100	94,4 %	94,1 %	96,6 %	94,7 %

Based on table 7, Linear Kernel achieves the highest value for precision, recall, and F1- Score. The results show that the model with linear kernel with hyperparameter C=0,1 has a better balance when compared to other kernels. It is important to consider the tradeoff between accuracy and model complexity when selecting kernels and hyperparameters. From the results, the linear kernel with C=0,1 has the highest accuracy with 98,9 %, followed by the poly kernel with C=100 with 96,7 % accuracy, and the rbf kernel with C=100 with 94,4 % accuracy.

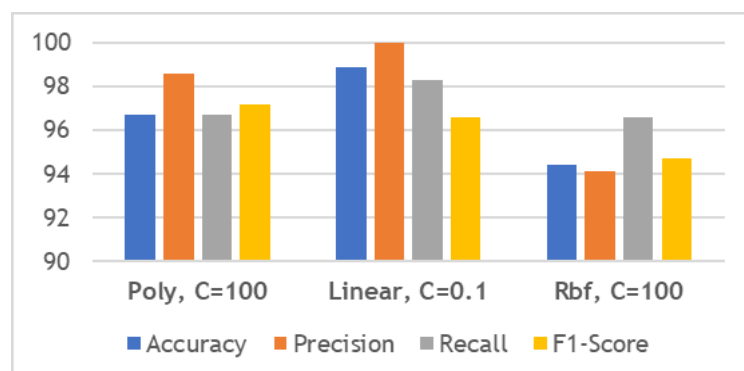


Figure 11. Performance Kernel Function with 10-Fold Validation for All Data

Considering the evaluation results and performance graphs in figure 11, the model with linear kernel ($C=0,1$) can be considered as the best choice in this case, as it has good and consistent performance on both datasets. This is because there is a good balance between bias and variance, then the simplicity of the model is more suitable for the linear kernel, then because the data is linearly distributed and does not require non-linear representation, and the tuning parameters are optimal in the linear kernel.

For ensure the classification results are consistent or not, testing with the Press's-Q is required, where the Press's-Q test can be written in equation (3) as follows:

$$\text{Press's-Q} = \frac{[N-nK]^2}{N(K-1)} \quad (3)$$

Where N is the total sample size, n is the number of correctly classified cases and K is the number of groups. The classification is accurate if the Press's-Q value is greater than the critical value of the Chi-Square distribution with an independent degree of one and the desired confidence level of 0,05.⁽²⁹⁾ Then, from the confusion matrix results in figure 10, we obtain the Press's-Q value as follows:

$$\text{Press's-Q} = \frac{[10-(10 \times 2)]^2}{10(2-1)} = 10 \quad (4)$$

If this Press's-Q value, i.e., Press's-Q = 10, is compared to the critical value of the Chi-square distribution which is 3,841, then the classification of parasitic and non-parasitic malaria images by using the SVM method is accurate, because the Press's-Q value is greater than the Chi-square distribution value.

DISCUSSION

The results of malaria parasite image feature selection for the malaria parasite *Plasmodium Falciparum* classification system based on blood smear images using the SVM approach are in image feature selection using the Principal Component Analysis (PCA) method. The PCA method, the normal malaria category with output 0 has PC1 and PC2 values that tend to be negative and more dispersed, with the average of each PC1 and PC 2 at output 0 being -123,85328961 and -1,36170437, respectively, while the parasitic malaria category with output 1 has greater variation in the PC1 and PC2 values. Output 1 has greater variation in both PC1 and PC2 components where the average of each PC1 and PC 2 at output 1 are respectively 123,85328961 and 1,36170437.

Classification analysis of the accuracy level generated from SVM approach are the average results of K-fold cross validation with all data and K-fold cross validation with data division for SVM models with three different kernel types, namely poly with hyperparameter $C = 100$, linear with hyperparameter $C = 0,1$, and RBFwith hyperparameter $C=-100$, for K-fold cross validation with all data, the average accuracy of each kernel is 96,7 %, 98,9 %, and 94,4 %. In the context of malaria parasite *Plasmodium Falciparum* classification research based on blood smear images, this research shows superiority compared to previous studies. The result of the accuracy analysis shows that the linear kernel in the SVM model with hyperparameter $C = 0,1$ achieves the highest accuracy result, which is 98,9 % in K-fold cross validation with all data. The strength of this research lies in the selection of the linear kernel and hyperparameter settings that resulted in good and consistent system performance on both datasets.

Previous research may not have specifically evaluated the performance of SVM models with a focus on specific kernel types and hyperparameters. The main contribution of this research lies in identifying the linear kernel as the best choice, providing a deeper understanding of the optimal parameters for malaria parasite classification. In addition, the use of K-fold cross validation with data sharing also provides additional insight into the generalizability of the model to new data. Thus, this study not only improves the accuracy of *Plasmodium falciparum* malaria parasite classification, but also provides new insights into the optimization of SVM models in this context. These positive results can serve as a basis for further development in the automated diagnosis of malaria through blood smear images.

CONCLUSIONS

Based on the results and discussion, this research demonstrated the effectiveness of an automated classification system for malaria parasites *Plasmodium Falciparum* based on blood smear images by using the Support Vector Machine (SVM) approach. The combination of Principal Component Analysis (PCA) for feature selection and SVM for classification proved to be highly effective, with the classification model achieving average accuracies of 96,7 %, 98,9 %, and 94,4 % across polynomial, linear, and radial basis function (RBF) kernels, respectively. The results indicate that SVM, particularly with a linear kernel, offers a reliable method for distinguishing between normal and parasitic malaria cases in blood smear images. This system has significant potential for aiding in the prompt and accurate diagnosis of malaria, contributing to improved management

and control of the disease, especially in high-risk regions by developing an accurate, automated system for diagnosing malaria parasites *Plasmodium Falciparum* using blood smear images, this study directly support SDG Target 3.3, which aims to end malaria and other communicable diseases. An effective diagnostic system allows for early detection and intervention, potentially reducing malaria-related mortality and morbidity in high-risk regions. Further research and development could refine this model for broader clinical use and integration into malaria diagnostic.

ACKNOWLEDGMENT

The authors would like to thank Dr. Budi Lestari, Drs., PGDip.Sc., M.Si., for providing useful comments, criticism and suggestions to improve the quality of this article.

BIBLIOGRAPHIC REFERENCES

1. Collins WE, Jeffery GM. *Plasmodium malariae* : Parasite and Disease. Clin Microbiol Rev [Internet]. 2007 Oct; 20(4):579-92. Available from: <https://doi.org/10.1128/cmr.00027-07>.
2. Varo R, Chaccour C, Bassat Q. Update on malaria. Medicina Clínica (English Edition) [Internet]. 2020; 155(9):395-402. Available from: <https://doi.org/10.1016/j.medcle.2020.05.024>.
3. WHO (World Health Organization). World malaria report 2023 [Internet]. World Health Organization; 2023. Available from: <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023>.
4. Elyazar IRF, Hay SI, Baird JK. Malaria Distribution, Prevalence, Drug Resistance and Control in Indonesia. In: Advances in Parasitology [Internet]. Elsevier; 2011. p. 41-175. Available from: <https://doi.org/10.1016/B978-0-12-385897-9.00002-1>.
5. Fong Amaris WM, Martinez C, Cortés-Cortés LJ, Suárez DR. Image features for quality analysis of thick blood smears employed in malaria diagnosis. Malar J [Internet]. 2022 Dec; 21(1):74. Available from: <https://doi.org/10.1186/s12936-022-04064-2>.
6. Hegde RB, Prasad K, Hebbar H, Singh BMK. Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. Biocybernetics and Biomedical Engineering [Internet]. 2019;39(2):382-92. Available from: <https://doi.org/10.1016/j.bbe.2019.01.005>.
7. Alam MM, Islam MT. Machine learning approach of automatic identification and counting of blood cells. Healthcare Technology Letters [Internet]. 2019 Aug;6(4):103-8. Available from: <https://doi.org/10.1049/htl.2018.5098>.
8. Ismael S, Kareem S, Almkhtar F. Medical image classification using different machine learning algorithms. AL-Rafidain Journal of Computer Sciences and Mathematics [Internet]. 2020;14(1):135-47. Available from: <https://doi.org/10.33899/csmj.2020.164682>.
9. Delgado-Ortet M, Molina A, Alférez S, Rodellar J, Merino A. A deep learning approach for segmentation of red blood cell images and malaria detection. Entropy [Internet]. 2020;22(6):657. Available from: <https://doi.org/10.3390/e22060657>.
10. Savkare SS, Narote SP. Automatic classification of normal and infected blood cells for parasitemia detection. Int J Comput Sci Net Sec [Internet]. 2011; 11:94-7. Available from: https://www.semanticscholar.org/paper/Automatic-Classification-of-Normal-and-Infected-for-Savkare/fc4073d02ab14b6c915bd1896526dcc91b9d8f45?utm_source=direct_link.
11. Zekar L, Sharman T. *Plasmodium falciparum* malaria. In: StatPearls [Internet]. StatPearls Publishing; 2023. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK555962/>.
12. Amiriebrahimabadi M, Rouhi Z, Mansouri N. A Comprehensive Survey of Multi-Level Thresholding Segmentation Methods for Image Processing. Arch Computat Methods Eng [Internet]. 2024 Aug; 31(6):3647-97. Available from: <https://link.springer.com/10.1007/s11831-024-10093-8>.
13. Bali A, Singh SN. A review on the strategies and techniques of image segmentation. In: 2015 Fifth international conference on advanced computing & communication technologies [Internet]. IEEE; 2015. p. 113-20. Available from: <https://ieeexplore.ieee.org/abstract/document/7079063/>.

14. Hassan NS, Abdulazeez AM, Zeebaree DQ, Hasan DA. Medical images breast cancer segmentation based on K-means clustering algorithm: a review. *Asian Journal of Research in Computer Science* [Internet]. 2021; 9(1):23-38. Available from: <http://archive.sdpublishers.com/id/eprint/131/>.
15. Pare S, Kumar A, Singh GK, Bajaj V. Image Segmentation Using Multilevel Thresholding: A Research Review. *Iran J Sci Technol Trans Electr Eng* [Internet]. 2020 Mar; 44(1):1-29. Available from: <http://link.springer.com/10.1007/s40998-019-00251-1>.
16. Kuşcu A, Erol H. Diagnosis of Breast Cancer by K-Mean Clustering and Otsu Thresholding Segmentation Methods. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi* [Internet]. 2022; 5(1):258-81. Available from: <https://dergipark.org.tr/en/pub/okufbed/article/994481>.
17. Bhatt SK, Srinivasan S, Prakash P. Brain Tumor Segmentation Pipeline Model Using U-Net Based Foundation Model. *Data and Metadata* [Internet]. 2023 Dec 30; 2:197-197. Available from: <https://dm.ageditor.ar/index.php/dm/article/view/108>.
18. Jimoh RG, Abisoye OA, Uthman MMB. Ensemble feed-forward neural network and support vector machine for prediction of multiclass malaria infection. *Journal of Information and Communication Technology* [Internet]. 2022; 21(1):117-48. Available from: <https://e-journal.uum.edu.my/index.php/jict/article/view/10958>.
19. Abdullah DM, Abdulazeez AM. Machine learning applications based on SVM classification a review. *Qubahan Academic Journal* [Internet]. 2021; 1(2):81-90. Available from: <https://journal.qubahan.com/index.php/qaj/article/view/50>.
20. Banu E, Geetha A. Hybrid Convolutional Neural Network with Whale Optimization Algorithm (HCNNWO) Based Plant Leaf Diseases Detection. *Data and Metadata* [Internet]. 2023 Dec 30;2:196-196. Available from: <https://dm.ageditor.ar/index.php/dm/article/view/109>.
21. Rachmad, A., Chamidah, N., & Rulaningtyas, R. Mycobacterium tuberculosis images classification based on combining of convolutional neural network and support vector machine. *Commun. Math. Biol. Neurosci.* [Internet]. 2020; Available from: <http://scik.org/index.php/cmbn/article/view/5035>.
22. Chamidah, N., Lestari, B., Saifudin, T., Rulaningtyas, R., Wardhani, P., Budiantara, I.N. (2024). Estimating the Number of Malaria Parasites on Blood Smears Microscopic Images Using Penalized Spline Nonparametric Poisson Regression. *Commun. Math. Biol. Neurosci.* [internet]. 2024; 2024, 60: 1-16. Available from: <https://scik.org/index.php/cmbn/article/view/8578>.
23. Muhimmah I, Lusiyan N. Identifikasi Stadium Plasmodium Ovale Penyebab Penyakit Malaria dari Apusan Darah Tipis dengan Sistem Berbantuan Komputer. *AUTOMATA* [Internet]. 2022;3(1). Available from: <https://journal.uui.ac.id/AUTOMATA/article/view/21904>.
24. Ramadhan NG, Khoirunnisa A. Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine. *Jurnal Media Informatika Budidarma* [Internet]. 2021;5(4):1580-4. Available from: <http://www.stmik-budidarma.ac.id/ejurnal/index.php/mib/article/view/3347>.
25. Andrew AM. An introduction to support vector machines and other kernel-based learning methods. *Kybernetes* [Internet]. 2001;30(1):103-15. Available from: <https://www.emerald.com/insight/content/doi/10.1108/k.2001.30.1.103.6/full/html>.
26. Durgesh KS, Lekha B. Data classification using support vector machine. *Journal of theoretical and applied information technology* [Internet]. 2010;12(1):1-7. Available from: http://jatit.org/volumes/twelfth_volume_1_2010.php.
27. Brereton RG, Lloyd GR. Support vector machines for classification and regression. *Analyst* [Internet]. 2010; 135(2):230-67. Available from: <https://pubs.rsc.org/en/content/articlehtml/2010/an/b918972f>.
28. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information processing & management* [Internet]. 2009; 45(4):427-37. Available from: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.

29. Rachmad A, Chamidah N, Rulaningtyas R. Mycobacterium tuberculosis identification based on colour feature extraction using expert system. Ann Biol [Internet]. 2020; 36:196-202. Available from: <https://agribioj.com/mycobacterium-tuberculosis-identification-based-on-colour-feature-extraction-using-expert-system/>.

FINANCING

The research was funded by Airlangga University, Indonesia, through the Airlangga University Mandate Research (Riset Mandat Universitas Airlangga) Grant with the contract number: 773/UN3.15/PT/2021.

CONFLICT OF INTEREST

The authors declare that no conflict of interest is associated with this research. The research process was conducted objectively and independently.

AUTHORSHIP CONTRIBUTION

Conceptualization: Riries Rulaningtyas, Nur Chamidah, Adam Anargya Mawardi.

Data curation: Adam Anargya Mawardi.

Formal analysis: Adam Anargya Mawardi, Naufal Ramadhan Al Akhwal Siregar.

Research: Nur Chamidah, Riries Rulaningtyas, I Nyoman Budiantara.

Methodology: Riries Rulaningtyas, Nur Chamidah, Toha Saifudin.

Project management: Nur Chamidah, Riries Rulaningtyas.

Resources: Riries Rulaningtyas, Nur Chamidah, Puspa Wardhani.

Software: Adam Anargya Mawardi, Toha Saifudin.

Supervision: Riries Rulaningtyas, Nur Chamidah, Puspa Wardhani, I Nyoman Budiantara.

Validation: Riries Rulaningtyas, Nur Chamidah, Toha Saifudin, I Nyoman Budiantara.

Display: Naufal Ramadhan Al Akhwal Siregar.

Drafting - original draft: Naufal Ramadhan Al Akhwal Siregar.

Writing - proofreading and editing: Nur Chamidah, Naufal Ramadhan Al Akhwal Siregar.