Bulletin of Applied Mathematics and Mathematics Education
Volume 4, Number 2, October 2024, pp. 141-150
https://dx.doi.org/10.12928/bamme.v4i2.11740
e-ISSN 2776-1029, p-ISSN 2776-1002

Identifying malaria disease through red-blood microscopic image with XGBoost and random forest methods

Rohmatul Fajriyah[1]*, Muhammad Muhajir[2], Ahmad Hussain Abdullah[3], Devina Gilar Ayu[4], Iqbal Fathur Rahman[5]

[1]Master Program in Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[2]Undergraduate Program in Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[3]Algoritma Data Science, Jakarta, Indonesia
[4]Astagraphia Information Technology, Jakarta, Indonesia
[5]Delta Dunia Makmur, Jakarta, Indonesia

*Corresponding e-mail: rfajriyah@uii.ac.id

ARTICLE INFO

Article History
Received 25 October 2024
Revised 10 December 2024
Accepted 12 December 2024

Keywords
Malaria
Microscopic blood image
Random forest
RShiny
XGBoost

How to cite this article:
Fajriyah, R., Muhajir, M., Abdullah, A. H., Ayu, D. G., & Rahman, I. F. (2024). Identifying malaria disease through red-blood microscopic image with XGBoost and random forest methods. *Bulletin of Applied Mathematics and Mathematics Education*, *4*(2), 141-152.

ABSTRACT

Blood cells that flow in the human body provide information to diagnose a disease. The information provided can be obtained through images of these blood cells using image processing techniques. Malaria is a very deadly disease and can affect everyone. Patients with malaria will experience anaemia because the red blood cells or erythrocytes are contaminated with plasmodium. This study offers an alternative solution to malaria disease identification through the image classification of red blood cells, by applying image processing and image classification methods with XGBoost and random forest. The research was conducted using the programming language in RStudio and Python. The accuracy of XGBoost and random forest methods were 71.26% and 77.58%, respectively. Therefore, the random forest provided a better optimal classification model with higher accuracy. The model is used to build an application which is R web-based, RShiny. In practice, this application can be used by health workers in classifying patients based on red blood cell images such that the health centre would be easier to manage the existing patients.

Introduction

Malaria is one of the infectious diseases that is a health problem in the world and in Indonesia in particular. The World Malaria Report (WMR) 2018 states that in 2017 an estimated 219 million cases of malaria occurred worldwide. Most of the malaria cases in 2017 were in the African Region (200 million or 92%), followed by the Southeast Asia Region (5%) and the Eastern Mediterranean Region (2%.) According to WMR 2018 in 2017 globally it was estimated that there were 435,000 deaths due to malaria. The most vulnerable group affected by malaria is children under the age of 5 years, where in 2017 this group accounted for 61% or as many as 266,000 of the deaths worldwide due to malaria (WHO, 2018).

The recent WMR 2023 stated that there were an estimated 249 million cases of malaria in the world and South-East Asia region had 5.2 million cases which contributes to 2% of malaria cases globally. In Indonesia, during 2021-2022, the cases and incidents are increasing. Somewhat interesting that in the world the cases are estimated to have decreased to 11.9% (WHO, 2023).

In the Presidential Regulation of the Republic of Indonesia number 59 of 2017 concerning the implementation of achieving sustainable development goals, the global targets by 2030 include ending malaria and fighting other infectious diseases (Kementerian Kesehatan, 2017) .This shows that in Indonesia malaria is still a serious problem that has not yet been handled properly.

Malaria is a disease caused by the bite of a female Anopheles mosquito that carries a protozoan parasite of the genus Plasmodium that can infect the red blood cells of the sufferer. There are several species of plasmodium that infect humans, namely Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae. Each species of Plasmodium causes a different malaria infection. Plasmodium vivax causes vivax/tertiana malaria, Plasmodium falciparum causes falciparum/tropical malaria, Plasmodium malariae causes malariae/quartana malaria, and Plasmodium ovale causes ovale malaria (Endah, 2020).

Plasmodium parasites survive by eating red blood cells where they nest such that the patience will experience anemia and other health problems. Hakim (Fitri et al., 2022) and Setyaningrum (Endah, 2020) stated that each species will require different breeding times. Therefore, species such as *Plasmodium malariae* and *Plasmodium ovale* are rarely found.

In the health sector, diagnosing malaria, is done by the anamnesis physical examination and laboratory support are carried out. Anamnesis examination is a diagnostic step that is carried out based on the examination of symptom complaints and patient history. Meanwhile, laboratory examination is a diagnosis that is carried out through examination with a microscope on thick and thin blood preparations to determine the presence or absence of malaria parasites, species and stages of plasmodium, and calculate their density (PDPERSI, 2019).

Along with the development of knowledge and technology, many other innovations are used by health workers in diagnosing malaria (Fitri et al., 2022), but from the various types of diagnostic developments currently available, microscopic examination is considered one of the best and used as a final determination of laboratory examination (Berzosa et al., 2018; Dayat & Banyal, 2018).

Microscopic examination has good sensitivity in identifying the types of parasites and their effects. However, the microscopic examination also requires a lot of trained microscopic experts and takes a relatively long time. Therefore, a fast and accurate method of diagnosis is needed for malaria patients. Preciseness and accuracy are helpful in determining the appropriate and correct treatment and management, as well as assisting in evaluating antimalarial treatment and resistance, which are very much needed by malaria patients (Dayat & Banyal, 2018).

The large number of malaria patients, limited resources, and the demand to provide laboratory results within a particular time such that malaria patients can be treated immediately and prevented from transmitting are our motivations in conducting the study to control malaria.

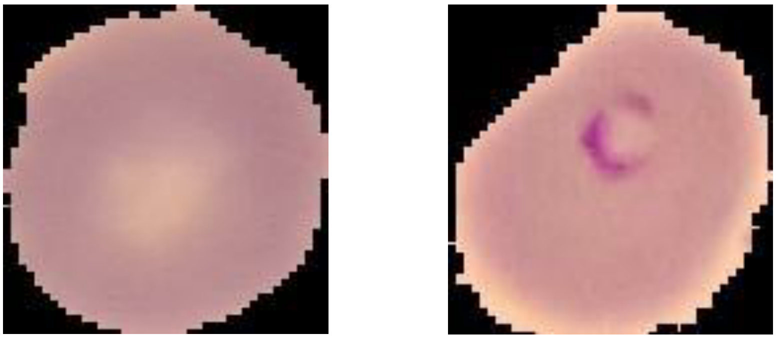
This study used a repository of microscopic images of the blood cells of 150 patients infected with malaria at the Chittagong Medical College Hospital, Bangladesh, (Dayat & Banyal, 2018). This microscopic image of human red blood cells, through specific image processing techniques, can be used to assist in the process of diagnosing the type of parasite that infects malaria patients (Gitta & Kilian, 2020; Madabhushi & Lee, 2016).

The image classification methods used in the red blood cell images of malaria sufferers are XGBoost and Random Forest. The optimal model is then poured into a prototype form of a web-based malaria patient identification application.

This paper is arranged systematically as follows. The first part is an introduction containing the motivation for conducting the research. The second part concerns the research methodology, population and sample, and matters related to data analysis methods. The third part presents the results of the research and its discussion, followed by the fourth part in the form of conclusions and remarks.

Method

The population and sample in this study are red blood cell images of malaria patients (Jain et al., 2020). There are 27,558 images with two categories, namely infected/parasitized cells and non-infected cells. The images example for the data are shown in Figure 1. It is a secondary data set. The data is taken from the https://lhncbc.nlm.nih.gov which is accessed on January 7, 2024.



(a) (b)

Figure 2. The image of uninfected blood cell (a) and the Image of infected/ parasitized blood cell (b).

In this study, to classify red blood cell images, the machine learning methods for classification XGBoost (Rahman, 2020; Shafila, 2020) and Random Forest (Rahman, 2020; Khoirunnisa & Ramadhan, 2023) methods were used. Data computing process is assisted by software R, RStudio and Python 3.6.2 with Keras and Tensorflow packages.

Xtreme Gradient Boosting (XGBoost)

The XGBoost is a machine learning method which is a combination of boosting and gradient descent created by Chen and Guestrin (2016) and Bridget (2021). Using the XGBoost, the end goal is to get function $F(x)$ hat with minimize the loss function $L(y, f(x))$, which is written as follows.

$$\hat{F} = \text{arg}min_f E_{x,y}\left[L\big(y, f(x)\big)\right] \quad (1)$$

In the training process, each iteration has to minimize the value of the loss function based on the initial function F_0 (x). In general, the gradient boosting algorithm has the following equation:

$$\{y_m, h_m\} = \text{arg}min \sum_{m-1}^{M} [\![L(y_i)]\!], f^{(m-1)} + y_m h_m(x_i)\Big) \quad (2)$$

In Chen and Guestrin (2016), it is stated that the XGBoost algorithm performs optimization 10 times faster than the implementation of other Gradient Boosting methods. The accuracy of the classification using XGBoost depends on the parameters used. The parameters on XGBoost can be seen in Table 1 (Rahman, 2020).

Table 1. Parameters in xtreme gradient boosting

Parameter	Details
Eta	Learning rate on training process
Gamma	Penalty parameter on regularization
Max_depth	The depth of a tree, the deeper the tree then it will be more complex
Min_child_weight	Penalty parameter on regularization
Subsample	Sample size used for the training process. For example, 0.5 means using half of the data randomly in creating a new tree
Colsample_bytree	Column sample size to create a new tree

Random Forest

According to (Segal, 2004) the random forest method is basically the development of the classification and regression tree (CART) method, through the implementation of bootstrap and feature selection. In this method, a number of trees grown to form a forest. The decision is then taken based on the mode among the yields of each tree in the forest (Kapwata & Gebreslasie, 2016).

The random forest algorithm according to (Tang et al., 2004), based on (Segal, 2004) is:

a. Perform bootstrap random sampling with replacement n, as many as n_{tree}
b. For each sample, grow a tree without pruning,
c. At each node, do sampling as many as m_{try}'s predictor. The rest will be on the Out of Bag (OOB).
d. Choose the best separator among the predictor variables (here will use the Gini index).
e. Repeat Steps 3 and 4 until there is no more best separator variable
f. Predict new data by aggregating predictions from n_{tree}.

Out of Bag

The variables that are not selected in the bootstrapping step are referred to as Out of Bag (OOB). This data is used as validation according to the tree. Classification errors are estimated through OOB errors. The calculation of the OOB error follows the procedure written in Jin et al. (2020), Breiman and Cutler (2003), Ummah (2019), and Tang et al. (2004), namely:

a. For each bootstrap sample, calculate the predicted OOB data using a tree that grows with the bootstrap sample;
b. Aggregate OOB predictions.
c. Calculate the error rate/proportion of misclassification (estimated OOB error rate)

OOB data is usually as much as 36% of the original data or a third of the number of trees formed. If is the original data, then the random forest's prediction is the composite of the predictions whenever becomes OOB data.

The OOB error depends on the correlation between trees and the strength of each tree in random fires (Breiman & Cutler, 2003; 2004). The higher the correlation (tree strength), the error will also increase. The number of explanatory variables (m) greatly affects the correlation and strength. The more, the correlation and tree strength will increase, and vice versa. Researchers need to determine the optimal m such that it will produce a small correlation but high strength. Therefore, a random forest with a small OOB error is obtained as described in Breiman and Cutler (2003; 2004), Breiman (2001), and Janitza and Hornung (2018). The suggested m as in Ummah (2019) are:

$$m = \{1/2\ \ |\sqrt{p}|, |\sqrt{p}|, 2|\sqrt{p}|\} \tag{3}$$

The importance variable will be obtained if a tree of at least 500 is used (Ummah, 2019). In the case of many explanatory variables, it is suggested that the number of trees is larger than the explanatory variables, such that the resulting importance variable is more consistent.

Mean Decrease Gini (MDG)

One measure of the variable importance of the explanatory variable in a random forest is the Mean Decrease Gini (MDG) (Breiman, 2001; Breiman & Cutler, 2004; Breiman & Cutler, 2003; Rahman, 2020). If you have explanatory variable with h= 1,2,...,p,then the MDG of explanatory variable X_h is calculated through

$$MDG_H = \frac{1}{K}\sum_t[d(h,y)I(h,t)] \quad (4)$$

where

$d(h,t)$: the magnitude of the decrease in the Gini index for the explanatory variable X_h at the vertices t $I\ (h,t) = \begin{cases} 1, & X_h \text{ sorted out the vertices } t \\ 0, & \text{others} \end{cases}$

k : number of trees in random forest (size of random forest).

Research Flowchart

In data analysis using XGBoost and random forest methods, the steps are taken as following.

a. Data Preparation
   - Ensure the data set will be used in the research
   - Checking and create the data training, testing and validation

b. Model Building
   - Create a function to convert an existing image into a numeric/pixel
   - Do the data wrangling for data training and testing
   - Choose cross-validation, k=2 fold
   - Do classification with XGBoost and random forest on train data
   - Validate the model by the data testing set to measure the model accuracy

Results and Discussion

In model building, we do the simulation to determine what bet hyperparameter for the XGBoost and Random Forest methods for the data at hand. We measure their performance after building the model based on the XGBoost and Random Foest methods. Prediction results from the testing data set based on the optimal random forest model can be seen in Figure 2.

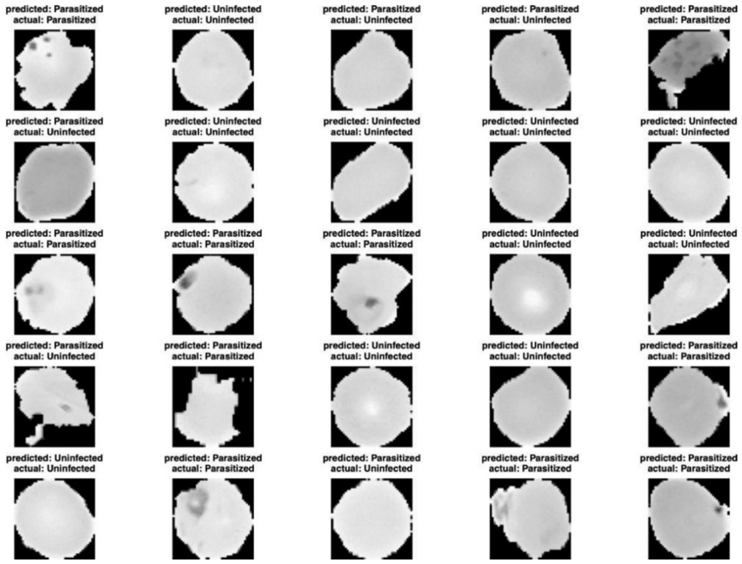


Figure 2. Visualization of prediction results from Random Forest data testing

Figure 2 shows that several prediction results are different from the original data. The results are understandable because the modeling using random forest, the obtained accuracy is 77.58%, see Table 2. Most uninfected data are predicted as infected (parasitized). These results were obtained because there are some blob/defects in the image and the image was darker. The blob and darker images lead the Random Forest method to classify them as parasitized.

Table 2. Accuracy of Random Forest and XGBoost data testing

	Random Forest	XGBoost
Accuracy	77.58%	71.26%
Recall	73.29%	71.55%
Specificity	73.29%	70.97%
Precision	75.49%	71.14%

Table 2 shows that the accuracy from the random forest is higher than the XGBoost. It means the model based on it performs better than the XGBoost. But still, the accuracy is lower than results from other researchers such as (Fuhad et al., 2020; Pan et al., 2018; Rajaraman et al., 2018, 2019; Ummah, 2019; Yang et al., 2020) It is understood that those researchers use deep learning methods to build the model where modified and pre-trained convolutional neural network (CNN) methods are applied some of which are preceded by feature selection, image acquisition, augmentation and segmentation.

Further, the best model in Table 2 is used to build the image analysis web-based application. The web-based application is built under the R programming and Python languages. It can be accessed at https://sbrc.shinyapps.io/jih-project/.

To use the application, first, the user needs to upload the blood cell image (PNG, JPEG) to the website. It can be done on the computer, laptop and or cell phone through the browse bar. Once the image is uploaded then the results will directly appear on the screen as Parasitized or Uninfected. Parasitized means that the patient is having malaria and Uninfected means that the patient is normal.

Suppose we have a blood cell image of two suspect malaria patients as in Figure 3. The web-based application gives the prediction of those images as (a) is uninfected and (b) is parasitized.

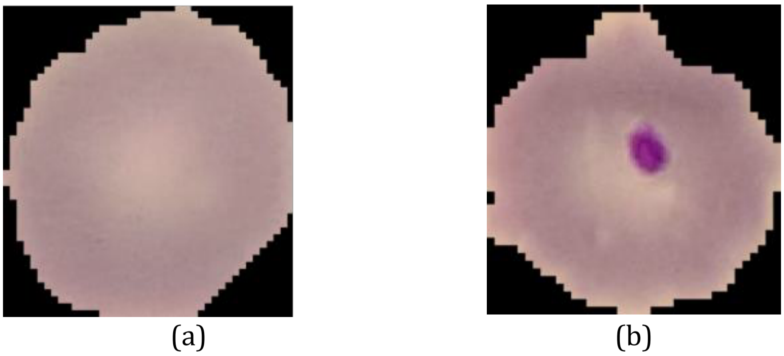


Figure 3. Microscopic blood image patients (a) and (b), source: https://lhncbc.nlm.nih.gov

Currently, the web-based application can accept any kind of image which the user can choose on their own devices. In the future, some limitations in the uploaded image will be applied. It should only be the microscopic blood image (single or non-single cell). Any other images will be directly identified as nonmicroscopic blood images by the application. This will make the application more specific and robust. We also will address the development of systems and apps as in WHO (2023), Bekele (2017), Das et al. (2013), and Yang et al. (2020).

Conclusion

Several conclusions were obtained based on the analysis that was carried out. First, we conclude that the Random Forest model provides better accuracy than XGBoost in identifying malaria through microscopic blood images to classify patient red blood cells. The accuracy is 77.58%. Second, to improve accuracy, methods related to image processing, pre-trained data, and classification methods can be implemented and adapted to the data available in Indonesia. In order to make the research benefit more evitable for health workers in Indonesia. Third, further research can be done to identify the type of malaria by considering the balanced and unbalanced aspects of the microscopic blood image data of the patient's red blood cells for each type of malaria, build a mobile phone-based application to make it easier for health workers in the public health units in every district in Indonesia and restrict the uploaded image to the microscopic blood image only.

Author Contributions

Rohmatul Fajriyah suggests the method, discussed with A. H. A. about R implementation to build the model and the app, and writes the manuscript. Muhammad Muhajir gives valuable input in the manuscript. Ahmad Hussain Abdullah implements the method in R and Python to build the model and the app. Devina Gilar Ayu writes the early manuscript. Iqbal Fathur Rahman implemented the method in R.

Acknowledgement

The authors would like to thank the DPPM Universitas Islam Indonesia and Jogja International Hospital Universitas Islam Indonesia for their funding.

Declaration of Computing Interest

We declare that we have no conflict of interest.

References

Bekele, A. (2017). Automatic detection of malaria parasite based on microscopic image analysis. *Doctoral Dissertation*. Addis Ababa University.

Berzosa, P., De Lucio, A., Romay-Barja, M., Herrador, Z., González, V., García, L., Fernández-Martínez, A., Santana-Morales, M., Ncogo, P., Valladares, B., Riloha, M., & Benito, A. (2018). Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea. *Malaria Journal*, *17*(1), 1–12. https://doi.org/10.1186/s12936-018-2481-4

Breiman, L. & Cutler, A. (2004). *Random Forests*. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Breiman, L. & Cutler, A. (2003). *Manual on setting up, using, and understanding random forest V4.0*. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Breiman, L. (2001). *Random Forest* (45th ed.). Springer. https://link.springer.com/article/10.1023/A:1010933404324

Bridget, O. N. (2021). Machine-learning techniques for malaria incidence and tuberculosis prediction. *Dissertation*. African University of Science and Technology. http://repository.aust.edu.ng/xmlui/handle/123456789/5096

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Das, D. K., Ghosh, M., Pal, M., Maiti, A. K., & Chakraborty, C. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, *45*, 97–106. https://doi.org/10.1016/j.micron.2012.11.002

Dayat, A. R., & Banyal, N. A. (2018). Penyebab penyakit malaria dalam sel darah merah manusia dengan menggunakan support vektor machine (SVM) di Kota Jayapura-Papua. *Jurnal ILKOM*, *10*(April), 28–32.

Endah, S. (2020). Mengenal malaria dan vektornya. *Bandarlampung, 53*(9).

Fitri, L. E., Widaningrum, T., Endharti, A. T., Prabowo, M. H., Winaris, N., & Nugraha, R. Y. B. (2022). Malaria diagnostic update: From conventional to advanced method. *Journal of Clinical Laboratory Analysis*, *36*(4), 1–14. https://doi.org/10.1002/jcla.24314

Fuhad, K. M. F., Tuba, J. F., Sarker, M. R. A., Momen, S., Mohammed, N., & Rahman, T. (2020). Detection from blood smear and its smartphone based application. *Diagnostics*, *10*(329).

Gitta, B., & Kilian, N. (2020). Diagnosis of malaria parasites Plasmodium sp. in endemic areas: Current strategies for an ancient disease. *BioEssays*, *42*(1), 1–12. https://doi.org/10.1002/bies.201900138

Jain, N., Chauhan, A., Tripathi, P., Moosa, S. Bin, Aggarwal, P., & Oznacar, B. (2020). Cell image analysis for malaria detection using deep convolutional network. *Intelligent Decision Technologies*, *14*(1), 55–65. https://doi.org/10.3233/IDT-190079

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *Plos One, 13*(8). https://doi.org/10.1371/journal.pone.0201904

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee turnover prediction based on random forests and survival analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35

Kapwata, T., & Gebreslasie, M. T. (2016). Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province South Africa. *Geospatial Health*, *11*(3), 251–262. https://doi.org/10.4081/gh.2016.434

Kementerian Kesehatan. (2017). Pedoman Teknis Pemeriksaan Parasit Malaria. *Buku Pedoman*, 1–78.

Khoirunnisa, A., & Ramadhan, N. G. (2023). Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy. *Jurnal Infotel*, *15*(4), 326–334. https://doi.org/10.20895/infotel.v15i4.1056

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, *33*, 170–175. https://doi.org/10.1016/j.media.2016.06.037

Pan, W. D., Dong, Y., & Wu, D. (2018). Classification of malaria-infected cells using deep convolutional neural networks. *Machine Learning - Advanced Techniques and Emerging Applications*. https://doi.org/10.5772/intechopen.72426

PDPERSI. (2019). *Buku Saku Penatalaksanaan Kasus Malaria*. http://www.pdpersi.co.id/kanalpersi/data/elibrary/bukusaku_malaria.pdf

Rahman, I. F. (2020). *Implementasi Metode SVM, MLP dan Xgboost pada Data Ekspresi Gen*. 1–79. https://dspace.uii.ac.id/handle/123456789/23679

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., & Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, *2018*(4), 1–17. https://doi.org/10.7717/peerj.4568

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest. *Center for Bioinformatics and Molecular Biostatistics*, *15*. https://escholarship.org/uc/item/35x3v9t4

Shafila, G. A. (2020). Implementasi metode extreme gradient boosting (XGBoost) untuk klasifikasi pada data bioinformatika. *Studi Kasus Penyakit Ebola*, GSE 122692, 1– 77.

Tang, G. H., Rabie, A. B. M., & Hägg, U. (2004). Indian hedgehog: A mechanotransduction mediator in condylar cartilage. *Journal of Dental Research*, *83*(5), 434–438. https://doi.org/10.1177/154405910408300516

Ummah, M. S. (2019). Properties of AdeABC and AdeIJK efflux systems of *Acinetobacter baumannii* compared with those of the AcrAB-TolC system of *Escherichia coli*. *Sustainability (Switzerland)*, *11*(1), 1–14.

WHO. (2023). *World Malaria Report*. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023

WHO. (2018). *World Malaria Report 2018*. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2018

Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., Maude, R. J., Jaeger, S., & Antani, S. (2020). Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. *IEEE Journal of Biomedical and Health Informatics*, *24*(5), 1427–1438. https://doi.org/10.1109/JBHI.2019.2939121

This page is intentionally left blank.

10.12928/bamme.v4i2.11740