JAIA – Journal Of Artificial Intelligence And Applications
Journal Homepage : http://jurnal.sar.ac.id/index.php/JAIA/

DETECTION OF MALARIA PARASITES IN HUMAN BLOOD CELLS USING CONVOLUTIONAL NEURAL NETWORK

Lusiana Efrizoni [1], Rais Amin [2], Ahmad Rizali [3]
[1]STMIK Amik Riau, lusiana@stmik-amik-riau.ac.id, Jl Purwodadi Indah KM 10, Pekanbaru, Indonesia
[2]STMIK Amik Riau, 2010031802119@sar.ac.id, Jl Purwodadi Indah KM 10, Pekanbaru, Indonesia
[3]STMIK Amik Riau, 1910031806004@sar.ac.id, Jl Purwodadi Indah KM 10, Pekanbaru, Indonesia

Article Info

Received : Jan 13, 2022
Revised : Mar 20, 2022
Accepted : Apr 10, 2022

Keywords :
Malaria
Data Science
Convolutional Neural Network
Multinomial logistic regression
Stochastic Gradient Descent
Nesterov momentum value

Abstract

Malaria is a blood disease caused by the Plasmodium parasite which is transmitted by the bite of the female Anopheles mosquito. The diagnosis of malaria is carried out by a microscopist through examination of human blood cells. Their level of accuracy depends on the quality of the tool, expertise in classifying and counting infected and uninfected parasite cells. The disadvantages of examining this way include the difficulty in making a diagnosis on a large scale and the poor quality of the results. The dataset used in model evaluation is a dataset developed by LHNVBC which contains 27,558 cell image data. The malaria dataset will be processed through data science processing using a Convolutional Neural Network with the ResNet architecture. The model will conduct training on the dataset and then the model will be able to recognize malaria parasites in human blood cells. The model will be trained by optimizing multinomial logistic regression using Stochastic Gradient Descent (SGD) and Nesterov momentum values. The results of training data validation accuracy from model training with 50 epochs were obtained at 96.23% and 97% after being tested on data testing.

1. Introduction

Malaria is still a public health problem that causes death, especially in high-risk groups, namely infants, toddlers, pregnant women, in addition to direct malaria cause anemia and can reduce work productivity. In 2018, an estimated 228 million cases of malaria occurred worldwide with a total number of deaths of 405 thousand. Most of the malaria cases in 2018 were in the African Region (213 million or 93%), followed by the Southeast Asia Region with 3.4% of cases and the Eastern Mediterranean Region with 2.1%.

Malaria case finding is based on clinical symptoms, through examination of blood preparations with laboratory confirmation using a microscope or Rapid Diagnostic Test (RDT).

The accuracy of blood smear examination is highly dependent on human expertise and can be affected by inter-observer variability and limited local resources for large-scale examinations in areas where malaria is endemic. Alternative techniques such as

Lusiana Efrizoni
Email: lusiana@stmik-amik-riau.ac.id

Polymerase Chain Reaction (PCR) and Rapid Diagnostic Test (RDT) have been used, however PCR analysis is limited in its performance and RDTs are less cost effective in large scale testing in areas where malaria is endemic.[1]

Several malaria datasets have been developed to assist in the examination and diagnosis of malaria. This study used a dataset (Lhnvbc.Nlm.Nih.Gov) which contains a repository of segmented cells from the Malaria Screener research activity. This dataset was developed by researchers at the Lister Hill National Center for Biomedical Communications (LHNVBC), which is part of the National Library of Medicine (NML).

This dataset was captured using a mobile application that has been developed, running on an Android smartphone, attached to a conventional light microscope. Giemsa-stained blood cells from 150 parasite-infected patients and blood cells from 50 healthy patients were collected and photographed at Chittagong Medical College Hospital, Bangladesh. The camera attached to the smartphone acquires a slide image for each microscopic field of view. The images are manually annotated by expert slide readers at the Mahidol-Oxford Tropical Medicine Research Unit in Bangkok, Thailand. NLM researchers apply a level-set based algorithm to detect and segment red blood cells. This dataset was created to reduce the burden on microscopists in areas where resources are limited and also to increase the accuracy of malaria diagnostics.

Data science processing is one of the processes that can describe predictive analytics protocols that are common to be interrogated on large and complex biomedical and health datasets. The process begins with identifying the problem, followed by determining data sources and meta-data, cleaning, alignment of data components, data preprocessing, model-based scientific inference, and ends with prediction, validation and dissemination of data, software, protocols and research findings.[2] From this description, data science processing can be the answer to be able to create an automatic diagnosis system from biomedical and health datasets, including malaria.

Automated diagnostic systems aim to perform this task of examining blood samples without human intervention and to provide an objective, reliable and efficient means of doing so. Automated diagnostic systems can be designed by understanding diagnostic expertise and representing them with special customized image processing, analysis and pattern recognition algorithms.from journals, conferences and books) about the topics discussed. The narrative builds on the general thing first and then gradually discusses the more specific one and finally emphasizes the benefits and objectives of the study which explains that this research really needs to be done.

This automatic diagnosis system can be designed using the process of implementing Machine Learning (ML) methods, accurate feature representation is the essence of success in achieving the desired results. The majority of image analysis-based automatic diagnostic systems use ML with hand-engineered features representation in decision making.[3]

However, this process requires special expertise in analyzing the variability of size, background, angle, and Region of Interest (ROI) in images. To overcome the challenges in designing hand-engineered features that capture variations in basic data, a Deep Learning (DL) model known as hierarchical learning is used which can be used to achieve significant success.[4]

Currently, many studies have applied DL and obtained promising results in various tasks of medical image analysis and comprehension. In *Evaluations of deep convolutional neural networks for*

automatic identification of malaria infected cells[5] and *Visualizing abnormalities in chest radiographs through salient network activations in Deep Learning*[6] prove that CNN has fewer yield parameters, less model complexity and computation time. And also proves that CNN has high accuracy and can also extract many layers of data input features automatically. Research [6], comparing CNN architectures (i.e AlexNet, VGG-16, ResNet, Xception and Densenet), and through a cross-validation study the results obtained that ResNet outperformed other CNN architectures in performance metrics for classification tasks, where the accuracy rate obtained was 95.59.

This study uses a CNN pre-taine-based model with ResNet architecture as an extractor feature in classifying parasitized and uninfected cells to assist in improving the malaria diagnosis system in human blood cells.

2. Research Methods

In this research process, a framework was created in the form of a schematic. The research methodology is used as a guide to determine the steps that must be carried out. These stages include:

2.1 Business Understanding

Several things are done at this stage such as understanding the needs and goals from a business point of view then interpreting knowledge in the form of defining problems in data and then determining plans and strategies to achieve goals.

2.2 Data Understanding

This stage begins with collecting data, describing the data, and evaluating the quality of the data.

2.3 Data Preparation

In this stage, namely building the final dataset which will later be trained using the model created. There are several things that will be done including making a work folder arrangement, dividing the dataset into training data, testing and validation, and also transforming and augmenting the data (Data Augmentation) to be used as input in the modeling and testing stages.

2.4 Modeling

This stage involves deep learning methods in determining and creating data training models. This study uses the Convolutional Neural Network (CNN) algorithm with the ResNet architecture.

2.5 Evaluation

This stage is carried out by looking at the performance level of the pattern generated by the algorithm. The parameters used to evaluate the comparison algorithm are the Confusion Matrix with the rules for accuracy, precision and recall. This value can be obtained by calculating:

$$Accuracy = \frac{True\ Positive}{Total\ Data} \times 100\%$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\%$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\% \quad (1)$$

2.6 Deployment

This stage is carried out by preparing reports and journal articles using the resulting model.

3. Results and Discussion

3.1 Problem Identification

Microscopists usually examine blood cells to diagnose malaria. However, their accuracy depends on the quality of the tools and expertise in classifying and counting parasitic and uninfected cells. Such an examination would be very

difficult in large-scale diagnostic processes and would result in poor quality.

The accuracy of blood smear examination is highly dependent on human expertise and can be affected by inter-observer variability and limited local resources for large-scale examinations in areas where malaria is endemic. Alternative techniques such as Polymerase Chain Reaction (PCR) and Rapid Diagnostic Test (RDT) have been used, but PCR analysis is limited in performance and RDT is less cost effective in large-scale testing in areas where malaria is endemic.

3.2 Data Collection

The author collects data and information by using literature studies and getting the results in the form of datasets taken from the National Library of Medicine which was developed by Stefan Jaeger in 2018 ftp://lhcftp.nlm.nih.gov/Open-Access-Datasets/Malaria/cell_images.zip which can be freely accessed for the benefit of application development and learning. This dataset will later be used as input data for the malaria parasite prediction system. The dataset consists of 27,558 cell images divided into two categories, namely parasitized and uninfected. Each category consists of 13,779 cell images.

3.3 Data Preparation

After the data has been successfully collected, then the data is continued to the data preparation stage.The first step in this stage is to divide the dataset into training, testing, and validation data.

In determining the ratio of the distribution of this dataset, for the ratio of the distribution of training data and data testing using the Simple Hold-Out Validation method [7], this method is the standard reference in the distribution of training and testing data with a ratio of 80:20 and for the ratio of the distribution of training data and validation data based on research [8] which compares the split values of 6:4, 7:3, 8:2 and 9:1. From this study it was concluded that the ratio of 9:1 produces a greater accuracy value than the other ratios.

Table 1. Distribution of data resulting from the division of the dataset

No	Folder Name	Category		Total
		Parasitized	Uninfected	
	Cell_images	13779	13779	27558
	Training	9955	9887	19842
	Validation	1098	1106	2204
	Testing	2786	2726	5512

Then, all the data that has been collected and divided into training, testing and validation folders are equated with all sizes, namely 64x64 pixels and the color category is "rgb" ("red", "green", "blue") and then given the category label "categorial". which is because each folder contains two class categories namely "Parasitized" and "Uninfected". Then the data in the training folder will be augmented using the Keras ImageDataGenerator module.

The training data is augmented by changing the image scale to 1/255 scale, and the image is rotated with range=20, and the zoom range=0.05, the width shift range=0.05, the height shift range=0.05, the shear range=0.05 and the image data is flipped horizontally. Then the data in the validation folder will also be augmented using the Keras ImageDataGenerator module by changing the image scale to 1/255. The data in this validation is not rotated and flipped because later it will be used as a validation test for the training data model that has been changed with validation data with the same image scale. The data in the testing folder is not augmented because it will later be used as test data for the model prediction stage against the original data.

3.4 Modeling

Modeling is done using the Convolutional Neural Network method with the ResNet network architecture.

This study uses an input image with a size of 64x64x3, the aim is to compare the accuracy value based on the size of the image. The network architecture in this study is explained as explained below:

1. The first convolution process uses a 3x3 kernel and a total of 64 filters. This convolution process is a combination process between two different matrices to produce a new matrix value. After the convolution process, an activation function is added, namely RELU (Retrified Linear Unit). This activation function aims to change negative values to zero (removing negative values in a convolution result matrix). The result of this convolution has a size of 64 x 64.

2. The first pooling process. This study uses maxpooling to get a new matrix value resulting from the pooling process. Based on the results of the pooling, it produces a new matrix of size 32x32 from the input of the first convolution result which measures 64x64. This process uses kernel pooling 3x3. The way maxpooling works is to take the maximum value based on the kernel shift as much as the stride value, which is 2.

3. The second covolution process is to continue the results of the first pooling process, namely with an input image matrix of 32x32 on 3 filters, namely 32, 32 and 128 filters and with a kernel size of 3x3. The second convolution process uses the RELU activation function. This process produces a 32x32 image.

4. The third covolution process is to continue the results of the second convolution process, namely with an input image matrix of 32x32 on 3 number of filters namely 64, 64 and 256 filters and with a kernel size of 4x4. The third convolution process uses the RELU activation function. This process produces a 16x16 image.

5. The fourth covolution process is to continue the results of the third convolution process, namely with an input image matrix of 16x16 on 3 number of filters namely 128, 128 and 512 filters and with a kernel size of 6x6. The fourth convolution process uses the RELU activation function. This process produces an 8x8 image.

6. The next process enters the second pooling process, this process is almost the same as the first pooling process, but this process uses average pooling with an 8x8 kernel size. This process produces the final output value of the matrix, which is a 1x1 image.

7. Flatten and fully connected. At this stage, only one hidden layer is used in the MLP (Multi Layer Perceptron) network. Flatten here changes the output pooling layer to a vector.

8. The final process is using the activation of the Softmax function. This function is specifically used in multinomial logistic regression classification methods and multiclass linear discriminant analysis.

Based on the description of the network architecture above, the architecture is used for the training process. So that from the training process a model of the architecture is obtained. The total parameters formed from the model are 2,164,174 neurons, and the parameters to be trained are 2,150,472 neurons.

3.5 Evaluation

In this process the model will be trained by optimizing multinomial logistic regression using Stochastic Gradient Descent (SGD)[4] and Nesterov momentum values[9]. The customized model is optimized for hyper-parameters with the grid search method[10].

In the training process this model will use the ResNet model architecture that has been created which is given the value input_shape = 64x64 pixels, 3 RGB image channels (“red”, “green”, “blue”), class 2

values where the dataset has 2 classes, kernel_size = (3 , 4 , 6) and with a regression value = 0.0005. Determining the value of the epoch for model training is 50 which is a repetition of the training process in one skip session 50 times in order to get the smallest error and produce good performance for the model made. Meanwhile, the loss category used is 'binary_crossentropy', and for the optimizer it uses 'SGD' with a learning rate = 1e-1 and a momentum value = 0.9.

By using 50 epoch iterations, the validation accuracy results are obtained as follows:

```
Epoch 50/50
620/620 [==============================] - 32s
522ms/step - loss: 0.1298 - accuracy:
0.973 - val_loss: 0.1392 - val_accuracy:
0.9623
```

Figure 1. Validation Accuracy

Based on the data above, it shows the results of model accuracy, it can be seen that the loss is 12% with a high accuracy of 97.3%. As for the validation data, the loss value is 13% and the accuracy is 96.23%. Which means it shows a model accuracy of 96.23%. With these results it can be proven that the modeling is successful and the system can very well distinguish between parasitized and uninfected image data.

After the modeling has been completed and the finalization stage, then the model is connected to the prediction program so that it can be run, used, and tested with data in the testing folder which contains 5,512 image data which is divided into 2 parasitized folders containing 2,726 image data and uninfected which contains 2,786 data picture. In the model testing phase the data in the images contained in the testing folder produces a very good accuracy value of 97%, where this result is very good which proves that the system can recognize and distinguish between parasitized and Uininfected images.

	precision	recall	f1-score	support
Parasitized	0.98	0.96	0.97	2726
Uninfected	0.96	0.98	0.97	2786
accuracy			0.97	5512
macro avg	0.97	0.97	0.97	5512
weighted avg	0.97	0.97	0.97	5512

Figure 2. Accuracy test results on data testing

From the data in the figure, it was found that the recognition accuracy in the Parasitized class was 97% of the 2,726 image data, and the same accuracy rate was 97% in the introduction of the Uninfected class. And the average yield of the two classes is 97%. This data proves that in the introduction of the Parasitized class, there were 2,644 image data that were successfully predicted correctly, and 82 image data that failed. Whereas in the Uninfected data class, 2,702 image data were successfully predicted correctly and 84 image data failed.

4. Conclusion

1. The CNN model that uses the ResNet network architecture in this study uses an input shape measuring 64 x 64 x 3, a filter size of 3 x 3, a total of 50 epochs. The data used for the model training process is 19,842 in the training folder and 2,204 data in the training folder. validation. Resulting in training and validation accuracy levels in detecting malaria parasites of 97.3% for training accuracy and 96.23% for validation accuracy. With these results it can be proven that the modeling is successful and the system can very well distinguish between parasitized and uninfected image data.

2. This study uses data testing of 5,512 images where per category there are 2,726 images in the Parasitized class and 2,786 images in the Uninfected class to be tested into the model that has been made. The testing results resulted in a new level of accuracy in detecting malaria parasites of 97%. With these results it can be proven that the modeling is successful and the system produces a very good level of

accuracy in categorizing and can very well distinguish and correctly label Parasitized and Uninfected image data.

5. Reference

[1] S. Rajaraman *et al.*, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 2018, no. 4, pp. 1–17, 2018, doi: 10.7717/peerj.4568.

[2] I. D. Dinov, *Data science and predictive analytics: Biomedical and health applications using R*. 2018.

[3] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Transl. Res.*, vol. 194, no. 2018, pp. 36–55, 2018, doi: 10.1016/j.trsl.2017.12.004.

[4] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[5] Y. Dong *et al.*, "Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells," *2017 IEEE EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2017*, pp. 101–104, 2017, doi: 10.1109/BHI.2017.7897215.

[6] R. Sivaramakrishnan, S. Antani, Z. Xue, S. Candemir, S. Jaeger, and G. R. Thoma, "Visualizing abnormalities in chest radiographs through salient network activations in Deep Learning," *2017 IEEE Life Sci. Conf. LSC 2017*, vol. 2018-Janua, pp. 71–74, 2017, doi: 10.1109/LSC.2017.8268146.

[7] F. Chollet, *Deep Learning with Python*. 2018.

[8] K. Akromunnisa, R. Hidayat, J. T. Informatika, and J. L. Adisucipto, "KLASIFIKASI DOKUMEN TUGAS AKHIR (SKRIPSI) MENGGUNAKAN K-NEAREST NEIGHBOR," vol. 4, no. 1, pp. 69–75, 2019.

[9] A. Botev, G. Lever, and D. Barber, "Nesterov's accelerated gradient and momentum as approximations to regularised update descent," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, no. 2, pp. 1899–1903, 2017, doi: 10.1109/IJCNN.2017.7966082.

[10] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.