Drug Resistant Prediction Based on Plasmodium Falciparum DNA-Barcoding using Bidirectional Long Short Term Memory Method

Lailil Muflikhah[1], Nashi Widodo[2], Novanto Yudistira[3], Achmad Ridok[4]
Faculty of Computer Science, Brawijaya University Malang, Indonesia[1, 3, 4]
Faculty of Mathematics and Natural Sciences, Brawijaya University Malang, Indonesia[2]

***Abstract*—Malaria disease mostly affects children and causes death every year. Multiple factors of the disease due to failure in treatment, including anti-malaria drug resistance. The resistance is caused by a decrease in the efficacy of the drug against Plasmodium parasites. Therefore, we proposed a computational approach using deep learning methods to predict anti-malarial drug resistance based on genetic variants of the Plasmodium falciparum through DNA barcoding. The DNA Barcode, organism identification from Plasmodium, is employed as data set for predicting the anti-malaria drug resistance. As a univariate amino acid sequence, it is transformed to numerical value data for building classifier model. It is constructed into a classifier model for prediction using Bidirectional Long Term-Short Memory (Bi-LSTM). This algorithm is extended from LSTM by two directions. In the first stage, the sequence is encoded into numerical data as input data for the method using sigmoid activation loss function. Then binary cross entropy is addressed to define the class, resistance or sensitivity. The final stage is applied by tuning hyper-parameter using Adaptive Moment Estimation optimizer to get the best performance. The experimental results show that Bi-LSTM as the proposed method achieves high performance for resistance prediction including precision, recall, and f1-score.**

***Keywords*—Drug resistant; plasmodium falciparum; Bi-LSTM; deep learning**

I. Introduction

Malaria is a type of infectious disease that is transmitted through mosquito bites or through blood transfusions. This type of disease exists in countries with tropical climates, including Indonesia. Various efforts have been made to reduce the number of deaths from this disease with intensive treatment. One of the causes of this high case is resistance to anti-malarial drugs against the plasmodium parasite. As it is known that the types of anti-malarial drugs are classified as antibiotics, so an appropriate regulation is needed. The development of efficient malaria vaccines is hampered by the inclusion of information about the huge number of genetic variants of this parasite, as many of these alleles avoid the allele-specific immunity elicited by vaccines. Potential vaccine candidate are merozoite proteins 1 and 2 (msp1 and msp2) and glutamate-rich protein from P. falciparum [1]. Resistance of the parasite Plasmodium falciparum to anti-malarial drugs is one of the causes of the high mortality rate in endemic areas.

One of the main causes is a mutation in the gene of the parasite, so that the drug is not right on the target region. Some genes are mutated in the anti-malarial drug Chloroquine (CQ) to the duplication of Mefloquine, it affected the Plasmodium to survive in the host's body. Several studies related to drug resistance were carried out both in-vivo and in-vitro to in-silico using the bioinformatics method approach to find the characteristics of these parasites. Molecular biology research was carried out in the wet lab to determine the level of polymorphism of the anti-malarial treatment target genes [2], [3]. Furthermore, based on a sufficiently high volume data set, research using a computational approach was conducted, how it is limited to image data (haploid), clinical data, and chemical data analysis to DNA sequence data (genetic variants) using machine learning algorithms [1], [4]–[8]. Therefore, we proposed to conduct research on study of genetic variant analysis against plasmodium parasite resistance in target genes, including the first transporter gene (Pfmsp-1) and second transporter (Pfmsp-1), as well as *Pfglurp* on the anti-malarial drug CQ with a computational approach using the deep learning method. Genetic variants are DNA sequence data from parasites that undergo mutations categorized as time series data. Long Term Sort Memory (LSTM) is a robust deep learning method for time series data. However, the right combination of hyper-parameters from this algorithm affects the performance. Then, this research aims to develop Bi-LSTM algorithm using tuning parameters to purpose a high level of performance.

II. Material and Method

A. Drug Resistance of Anti-Malaria

Treatment of malaria is through drug therapy which aims to eliminate the plasmodium parasite in the host's body. However, the resistance of the parasite *plasmodium* falciparum to anti-folate resulted in the sequential acquisition of *mutations* in the target gene as shown in Fig. 1. The high mutation rate of this type of Plasmodium parasite species causes resistance, so a combination of anti-malarial drugs is carried out. Including efforts for possible immunity of malaria parasites to anti-malarial drugs, WHO has advised the use of an approved combination of artemisinin, Artemisinin Combination Therapy (ACT) [9].

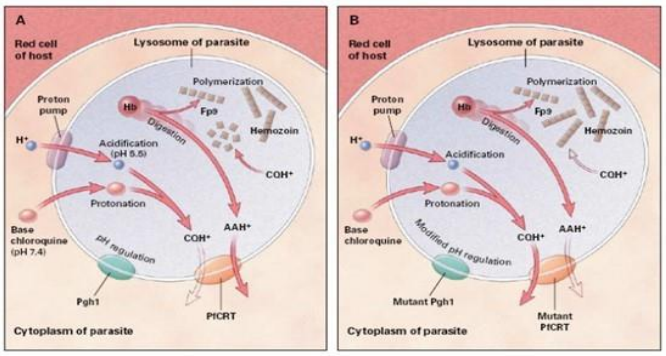

Fig. 1. Illustration of resistance of plasmodium parasite to anti-malarialdrug CQ [9].

B. DNA Barcoding

DNA barcoding is a strategy to recognize and to give auto-validation for living beings. It has a specific region, and the length of the sequence is between 300 and 400 bps. A method for automatically identifying and validating living things is DNA barcoding. The sequence has a specific region and a length between 300 and 400 bps. According to Li, et al. (2018), the barcoding-based protein measuring approach converts protein signals into barcoded oligo probes and amplifies the signals with nucleic acid amplification to achieve ultrasensitive detection down to the single molecule level [10]. With this technique, it is possible to identify rare tumor cells or cancer cells with low abundance surface markers [11]. The Weissleder group developed photo-cleavable DNA- barcoded antibodies that are specifically capable of recognizing multiplexed cell biomarkers [10]. After that, DNA has been cleaved by light (less than 365 nm) and released into solution; gel electrophoresis can be used to analyze it. They also came up with an amplification-free method to profile more than 90 proteins in single cells using DNA-barcoded antibodies and Nano String's fluorescent readout [12] to study the drug response pathway and inter- and intra-tumor heterogeneity in clinical samples [12], [13]. DNA barcoding can be used to study the resistance of Plasmodium parasites to various anti-malaria drugs. Here are a few examples:

1) Chloroquine was once the most widely used anti-malaria drug, but its effectiveness has been greatly reduced by the emergence of chloroquine-resistant Plasmodium falciparum parasites. DNA barcoding can be used to identify mutations in the parasite's Pfcrt gene that are associated to chloroquine resistance.

2) Artemisinin and its derivatives are currently the most effective anti-malaria drugs, but artemisinin-resistant parasites have emerged in Southeast Asia. DNA barcoding can be used to identify mutations in the parasite's K13 gene that are associated with artemisinin resistance.

3) Sulfoxide-pyrimethamine is a combination drug that is used to treat malaria and is also used for intermittent preventive treatment in pregnant women. Resistance to sulfoxide-pyrimethamine is common in many parts of the world. DNA barcoding can be used to identify mutations in the parasite's *Dhfr* and *Dhps* genes that are associated with sulfoxide-pyrimethamine resistance.

C. Machine Learning

Machine learning is a way of learning from data that has been set up in a way that is supervised or directed. This means that a set of instructions is given to the machine learning algorithm, telling it what to look for and how to interpret it. Unsupervised learning is when the machine learning algorithm is left to its own devices, without any specific instructions. In machine learning, we use supervised learning to learn how to map input (X) to output (Y). This is done by having a set of input values and a set of output values and using algorithms to find a function that can map the input to the output. We want to be able to predict the output variable for given input data. In supervised learning, we give the machine lots of data that tells it what to do. With unsupervised learning, we just give the machine data [14].

The unsupervised learning method is used to model the underlying structure or distribution in the data so that it can be studied more closely [15]. In this research, we propose and evaluate the supervised method, such as KNN, Naïve Bayes, SVM, Random Forest, and advanced machine learning (deep learning) using LSTM and Bidirectional LSTM (Bi-LSTM).

D. Deep Learning

Deep learning is an enhanced artificial neural network (ANN) method with multiple-hidden layers. In principle, deep learning is a neural network with three or more layers of ANN, enabling it to learn and adapt to large amounts of data and to solve various problems that are difficult to solve by other machine learning algorithms [16]. As illustration, the architecture of deep learning is shown in Fig. 2.

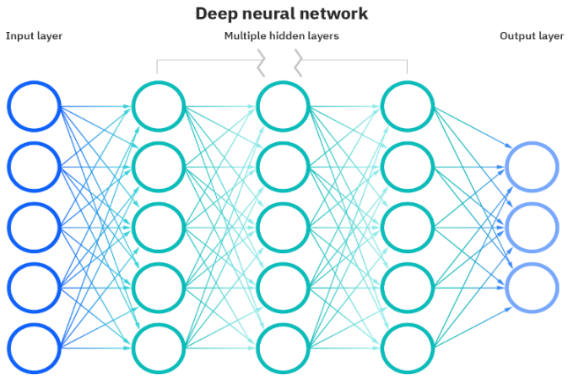

Fig. 2. Deep learning architecture [16].

Deep learning consists of several artificial neural networks that relate to each other. Some types of algorithms in deep learning including Convolution Neural Network (CNN), Recurrent Neural Network (RNN), and Self-Organizing Map (SOM),

1) Recurrent neural network: Recurrent Neural Network (RNN) is a neural network model which uses loops in its internal memory to solve sequential data problems [17]. The architecture of the RNN can be seen in Fig. 3.

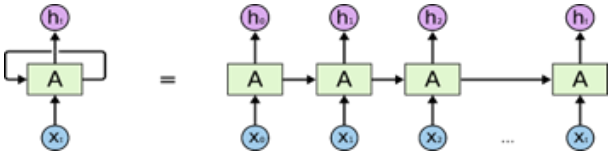

Fig. 3. The architecture of recurrent neural network.

Based on this architecture, each hidden layer of the RNN will receive input to the form of a vector denoted as Equation 1.

$$h_t = \sigma(W_{xh} + W_{hh}h_{t-1} + b_h) \quad (1)$$

where, W_{xh} is the weight matrix from the input to the hidden layer; W_{hh} is the weight matrix between two hidden states; b_h is a bias vector of hidden layer; σ_h is an activation function for getting hidden state. After getting the value h_t, then thenext process is to find Y_T as in Equation 2.

$$Y_t = \alpha_h(W_{hy}h_t + b_y) \quad (2)$$

where W_{hy} is weight matrix of hidden layer into output layer; by is bias vector of output layer; σyis activation function of output layer.

2) Long Short-Term Memory (LSTM): Long Short-Term Memory is one of the Recurrent Neural Network (RNN) architectures created to handle the problem of gradients that disappear during the back-propagation process [17]. LSTMs are special types of neural networks that can remember things for a long time. Each memory cell stores information by using an input gate (which lets in new information), a forget gate (which helps to forget old information), and an output gate (which sends out the stored information). The forget gate helps to forget things from the past, the input gate helps you transfer information into the cell, and the output gate helps you create new memories in the long term [18]. The input to the LSTM is a sequence of numbers$(x_1,x_2,\ldots,x_t)$. The output is a sequence of numbers $(y_1, y_2,\ldots,y_n)$. To find out how many forget gates there are, we use Equation 3. This equation tells the frequency of output will change in a particular gate.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (3)$$

3) Bidirectional Long Short Term-Memory (Bi-LSTM): Bidirectional Long Short-Term Memory is the LSTM method with two directions- a forward layer to remember information and a backward layer to remember how information was learned - to make it more efficient [19]. LSTMs are better at handling situations where the gradient of a neural network's activation changes slowly over time, like in language processing. However, LSTMs only work with information from the past [20]. To fix this problem, the LSTM architecture needs to be changed so that it includes a forward LSTM layer and a backward LSTM layer. This will help the machine learn what happened in the past and make predictions for the future. As illustration of the architecture is shown in Fig. 4.

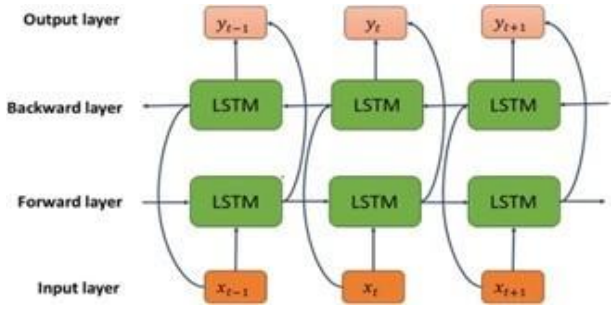

Fig. 4. Architecture of Bi-LSTM.

According to Cui et al.(2022), the forward layer's output (h_f) is calculated using the forward input starting at 1 – T, while the backward layer's output (h_b) is calculated using the inverse of the backward layer's input, namely T – 1 [21]. The gate that contains a Forward layer and a retrogressive layer can be refreshed involving a similar condition as a conventional LSTM. In the hidden state section of BiLSTM, the t-the data contains both the hidden state forward layer and the hidden state backward layer. This is the primary distinction. The equation can be used to combine BiLSTM hidden state.

$$h_t = f(h_{tf}, h_{th}) \quad (4)$$

where htf is the hidden state of the forward layer, htb is the hidden state of the backward layer, and f is a function to combine the output of the hidden state forward layer and backward layer. The function f can be a concatenating function, a summation function, an average function, or a multiplication function.

E. Optimizer

An optimizer is an algorithm that helps to find the lowest possible global value for the convergence of a loss function in a neural network [17]. An optimizer can be used to find the best values for the model's parameters [22]. This helps to improve the model's accuracy. The optimizer can also help to reduce the model's error value. Adaptive learning methods like Adam help us estimate how well we're doing during a learning task. Adam is an optimization of an existing, effective stochastic algorithm that doesn't require a lot of memory. Its streamlining agent calculation is planned by joining the strengths of the other two enhancers, Adagrad and RMSprop.

1) Adaptive Moment Estimation (Adam): Adaptive Moment Estimation (Adam) is one of the methods adaptive learning [18]. Adam is an optimization of an efficient stochastic algorithm that only requires a first order gradient where the memory required is very little [23]. Adam's optimizer algorithm is designed by combining the advantage of the other two optimizers, Adagrad, which has performance pretty good with sparse gradients and RMSprop which works well at times solve non-stationary problems. Updating the weights on the optimizer Adam can be seen in Equation 5 to Equation 7.

$$V_t = \beta_t * V_{t-1} - (1 - \beta_1) * g_t \quad (5)$$

$$s_t = \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2 \quad (6)$$

$$W_{new} = W_{old} - \propto \frac{v_t}{\sqrt{s_t + \varepsilon}} * g_t \quad (7)$$

2) Stochastic Gradient Descent (SGD): Stochastic gradient descent is a way of finding the smallest possible error in a calculation. It begins by making small changes to the current value and then checking to see if the error has decreased [22]. If it has, the algorithm keeps making smaller changes until the error decreases no further. The SGD optimizer updates the weight of a model as in Equation 8.

$$w_{new} = w_{old} - (\alpha * g_t) \quad (8)$$

where gt is the gradient of the t time step, w is weight, and α is the learning rate.

III. The Proposed Research Method

In general, this research is applied using Bi-LSTM as shown in Fig. 5. It was started by collecting DNA barcode data from several falciparum parasite DNA barcode sequences for anti-malarial drug resistance at Malaria GEN sites, including Artemisinin, Piperaquine, DHA-PPQ, Chloroquine, Pyrimethamine, Sulfadoxine, and S-P. The various drugs are comparable to the anti-malaria drug data set with any possibility misuse in regulation by patients. The drugs are composed in the type of antibiotic. The resistance prediction study includes preprocessing data, determining input data, developing the RNN method (Bi-LSTM algorithm), tuning parameters to optimize the Bi-LSTM method, and evaluating performance through measuring accuracy, precision, recall, and f1-score**.**

We used seven different datasets of drug with two classes, namely resistant and sensitive as shown in Table I. The data set is taken from MalariaGEN Plasmodium Falciparum Community Project which analyzed 7,113 parasite samples obtained at 73 different sites in Africa, Asia, America, and Oceania to generate genotyping data.

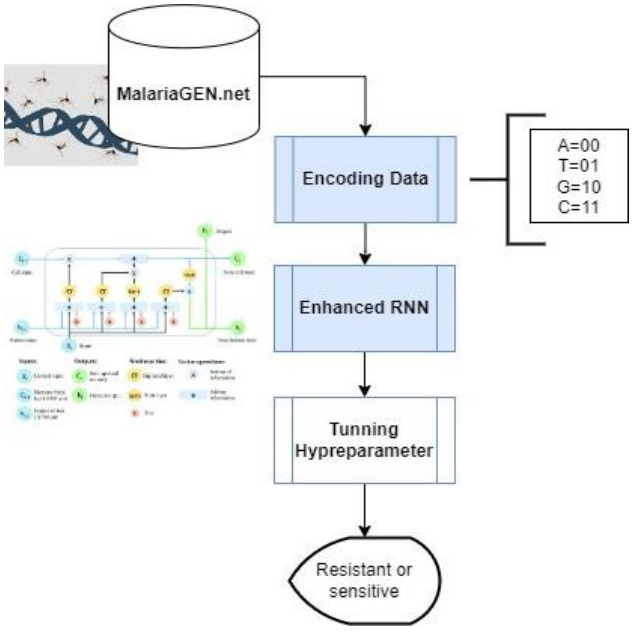

Fig. 5. Concept diagram of drug-anti-malaria resistance prediction.

TABLE I. Data Sets

Drugname	Resistant	Sensitive
Artemisinin	3176	3787
Piperaquine	2141	1297
DHA-PPQ	1889	4154
Chloroquine	6402	430
Pyrimethamine	7378	120
Sulfoxidess	5967	1412
S-P	4553	1729

By parsing the dataset, we got the DNA-barcode sequence as shown in Fig. 6. The barcode sequence represents more than 100 nucleotides in each sample of plasmodium.

SampleId	Species	GenBarcode	Artemisinin
RCN03189	Pf	AAXXXXXATXAACGCGXATXAGGAGCGGCCAACCTCATTXCACXCATGCAAAGGTGGGXCTAAGTAACGTAACXCXGXCTCXCAXTAXXCXXXCAACXAAT	Resistant
RCN03192	Pf	GGXXGACXXXAGTGTXGXAXTGTGACACTCAATCATATTXATTXTXTAXXAXAGTGGGXXTXXGXAXCAXCXGXAXGXCXCXXAXTGXXXCTXXXXTGAAT	Sensitive
RCN00121	Pf	GGXCANTGTXAANCNGCNTNATTNGCXCCCTNTCATGTCGCNTTNGNGCAXAAXNNNGXACXNACGCNNAAXGXAGAGTCAGNGACNATCTCGCTACNXNT	Sensitive
RCN00122	Pf	AGXTAATACXGATXTCCGAAATTAATXXCCTGTCACATCCCAXXCATAXXXAAXTGGAXCCXAGTXXXGTCXCXXGACTCAXAGCTAXCACCACTTCAXGT	Sensitive
RCN00123	Pf	AATTGATATXGATCTGXAXAATGGGCXXCTAACCTTAATGXAXTCATAAGXGGXCAGGXATAAGCGXXGTCXGXCGAGTCCGAAXCGATATCGCTATAXGC	Sensitive
RCN00124	Pf	GGGCGGCATXGGTCCGCATCAGXAGTGGCTAGTTATGATGCATTCAXGCAXAGXCGGAXXCTGATGCCGACAGCAGGCCTCGAGACAACACCGCTACGAGT	Sensitive
RCN00125	Pf	XXXXXXXXXXXXXXXXXXXXACATTGATGXTTTGTCACAACGCAXXTATAAGXAGXTATAXACXGGTGCXGACXCXCGGCTCATCGACAACATCGCTACGXGT	Sensitive
RCN00126	Pf	GAXTGGXGXXAACCCGTAXAAXTAATXCCTTATCTXAXTCCTXTTATAAGXAGXCGTGXATTGACGATGTCXCXAGGGCTCGAAATATTCCCGCTTCAAGT	Sensitive
RCN00127	Pf/Pv	GGXCGGTGCXGACXCGCGTCAGGAGCXCCCAATCTTGTCCCATTCGTAXXXAAXTATGXACXAACACTATCXCXXGGCTCCGAAACAATACCACTACAXGT	Sensitive

Fig. 6. The sample of DNA barcode sequences.

The stages of the proposed method are shown in Fig. 7. The design of an algorithm for predicting anti-malarial drug resistance against the plasmodium falciparum parasite is carried out by incorporating the DNA barcode sequence dataset into the model architecture. Before data sequence is entered, the dataset is encoded first in the stage preprocessing by splitting the dataset into training data and test data and perform encoding. The next step is the forming process architecture using the LSTM algorithm. Then, it is performed training model as well as the model validation process that has been formed to be carried out. After the model was trained, it was tested to evaluate the performance of the model in classifying drug resistance.

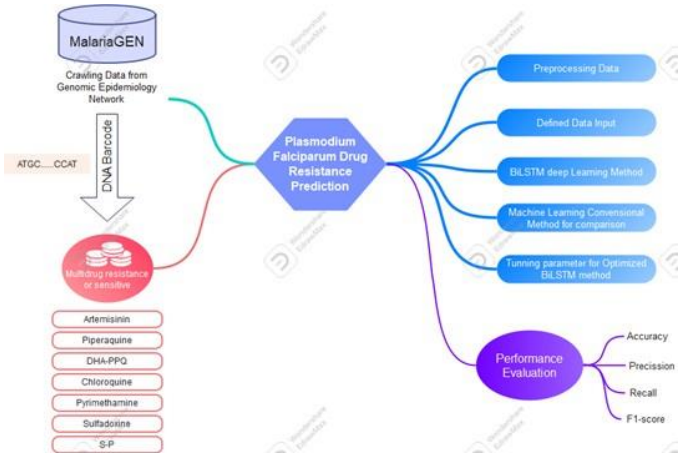

Fig. 7. The Stages of proposed method.

IV. Result and Discussion

Anti-malaria drug is a kind of antibiotic drug. Patient non-compliance in the use of anti-biotic drugs has an impact on resistance to pathogens in infecting a person's body. Many

studies of drug resistant for anti-malaria were conducted and various attempts have been made to treat this disease [11]. DNA barcoding can be used to study the resistance of anti-malaria drugs. DNA barcoding is a technique that involves sequencing a short, standardized region of DNA from a particular organism. In the case of malaria, DNA barcoding can be used to recognize the specific species of the malaria parasite present in a patient's blood [24].

The genetic sequence of the malaria parasite: It provided information about its susceptibility to different anti- malaria drugs. Mutations in the parasite's DNA can result in resistance to certain drugs, and DNA barcoding can be used detect these mutations. In addition to identifying resistance to specific drugs, DNA barcoding can also be used to monitor the spread of drug-resistant malaria parasites. By sequencing the DNA of malaria parasites from different regions, researchers can track the movement of drug-resistant parasites and identify areas where resistance is becoming more prevalent. Overall, DNA barcoding is a valuable tool for studying the resistance of anti-malaria drugs and developing strategies to combat drug-resistant malaria. The attribute values distribution of data set is plotted into boxplot as shown in Fig. 8. The graph shows that the variation in attribute values is distributed in the range between 1 and 7 and the first to third quartiles are described. The presence of a vertical line through this plotted box corresponds to the median distribution of the partial data. The boxplot is illustration of statistical analysis for data distribution [25]. They are range value of DNA-barcode sequences after encoding in numerical values.

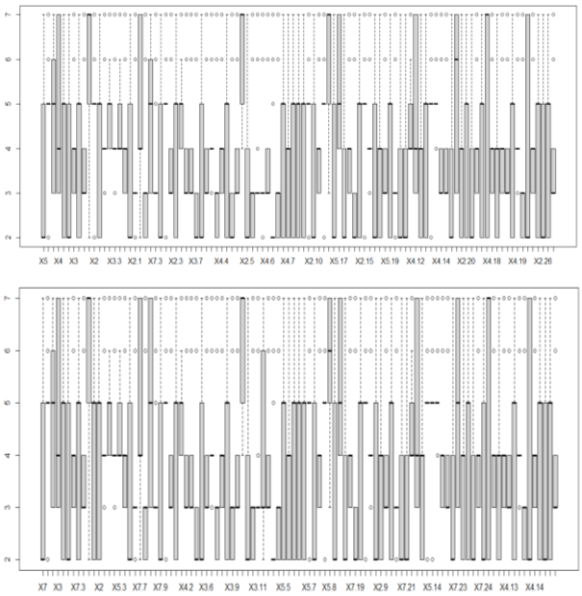

Fig. 8. Boxplot of encoding barcode DNA sequence dataset distribution for training and testing data.

This research is developed using the Python programming language, and Google Collaboratory as a Python text editor. We assessed the grouping model by ascertaining accuracy, precision, recall, and the F1-score to determine the quality of result. The number of detections is True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) was calculated as follows to calculate these measurements. The number of tests (drug resistance accurately) is TP. The number of tests incorrectly identified a drug resistance is called FP. The number of tests that were incorrectly classified as sensitive is referred to in FN. Then again, TN is the degree of tests that were precisely settled to be solid. Additionally, a five k-fold cross-validation method was utilized to ensure the quality of the classification model for detection.

As a measure of correctness, accuracy is defined as the ratio of the number of corrected forecasts to the total number of forecasts, as shown in Equation (9). By then, the pace of positive conjectures that are exact is referred to as precision as in Equation (10). As shown in Equation (11), Recall is the total number of positive cases that the classifier correctly anticipated based on all the information. Sometimes it's referred to as affect ability. The F1-score is the final performance metric. A degree that combines review and precision could be the score as shown in Equation (12) [26].

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$precision = \frac{TP}{TP+FN} \quad (10)$$

$$recall = \frac{TN}{TN+FP} \quad (11)$$

$$f1-score = 2 \times \frac{precision \times recall}{precision+recall} \quad (12)$$

The representative machine learning methods including *K-Nearest Neighbor*, *Naive Bayes*, *SVM*, *Decision Tree*, and the deep learning methods including, *RNN*, *LSTM*, and *Bi-LSTM* are used to evaluate the performance measures including *accuracy*, *precision*, *recall*, and f 1-*score*. The performance results for each algorithm described into the heat map in Fig. 9. All methods have high performances evaluation for two kinds of anti-malaria, such as Dha-Ppq, and Piperaquine. However, the conventional machine learning methods have low in unbalanced class of data sets. The distribution of class in each data set can be seen in Fig. 10.

Based on the experimental results, we can see that most of the classic machines learning models have low performance rates in unbalanced class distribution as shown in Table II and Table III. A classifier whose score is less than 0.5 has a high proportion of false negatives. It can be caused by an imbalanced class or the number data in each class is imbalanced. It means that classic machine learning model can't overcome imbalanced class problem like deep learning. The over-sampling or under-sampling strategy, we attempt to change the amount of the minor class to obtain a balanced proportion for both classes. A potential solution may help to resolve the imbalanced class problem for the traditional machine learning model Random Over-sampling, Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Random Under-sampling, etc. are some over-sampling and under-sampling techniques that are frequently utilized. Another solution that we can try is to give higher weight to the minor class. A class has higher weight than the other class [27]. By adjusting the cost function of the model, class weighting ensures the misclassified observations. The minority class tends to decide the majority

class. By rebalancing the class distribution, this strategy can assist in enhancing the model's accuracy.

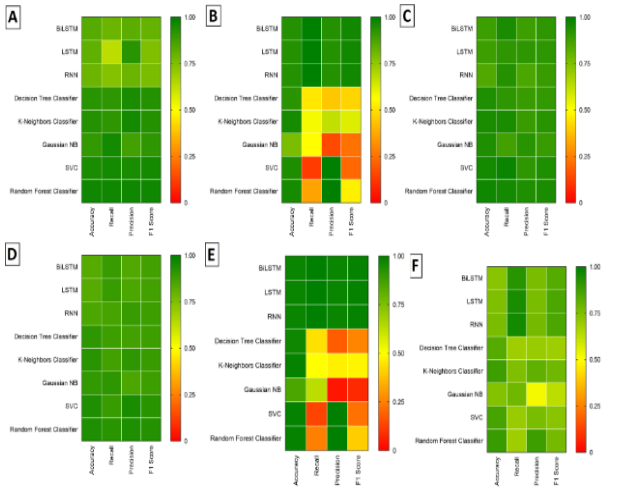

Fig. 9. Comparison of performance evaluation for various drugs including A. Artemisin, B. Chloroquine, C. Dha-Ppq, D. Piperaquine, E. Pyrimithamine, F. S-P,G.Sulfoxides.

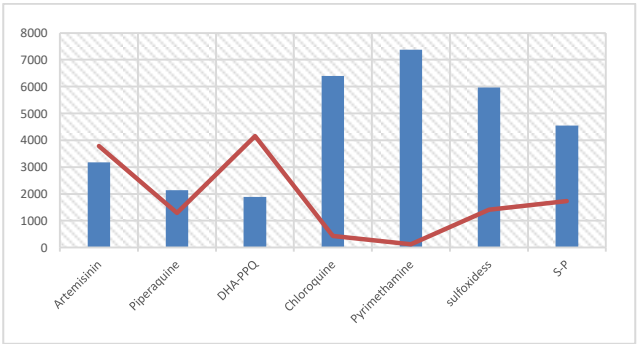

Fig. 10. Distribution of class in each data set.

TABLE II. PERFORMANCE RATE OF CHLOROQUINE

Model	Accuracy	Recall	Precision	F1 Score
BiLSTM	**0.94**	**1.00**	**0.94**	**0.97**
LSTM	0.94	1.00	0.94	0.97
RNN	0.94	1.00	0.94	0.97
DecisionTree	0.93	0.51	0.44	0.47
Kneighbors	0.96	0.58	0.67	0.62
GaussianNB	0.79	0.56	0.16	0.25
SVC	0.95	0.13	1.00	0.23
RandomForest	0.96	0.36	1.00	0.53

TABLE III. PERFORMANCE RATE OF PYRIMETHAMINE

Model	Accuracy	Recall	Precision	F1 Score
BiLSTM	**0.98**	**1.00**	**0.98**	**0.99**
LSTM	0.98	1.00	0.98	0.99
RNN	0.98	1.00	0.98	0.99
DecisionTree	0.97	0.44	0.18	0.26
Kneighbors	0.99	0.50	0.47	0.48
GaussianNB	0.85	0.63	0.04	0.08
SVC	0.99	0.13	1.00	0.22
RandomForest	0.99	0.25	1.00	0.40

TABLE IV. PERFORMANCE RATE OF ARTEMISININ

Model	Accuracy	Recall	Precision	F1 Score
BiLSTM	**0.83**	**0.79**	**0.82**	**0.80**
LSTM	0.81	0.63	0.92	0.75
RNN	0.79	0.74	0.79	0.76
DecisionTree	0.92	0.91	0.95	0.93
KNeighbors	0.93	0.91	0.96	0.93
GaussianNB	0.90	0.96	0.87	0.91
SVC	0.95	0.96	0.95	0.96
RandomForest	0.97	0.97	0.97	0.97

Furthermore, the high-performance scores were obtained for the Bi-LSTM algorithm under various data conditions (balanced or unbalanced) when compared to the other deep learning methods. In Table IV, the performance of Bi-LSTM for Artemisinin data set with balanced class distribution, can provide consistent and reliable results. Even though the conventional machine learning algorithms achieved higher performance than Bi-LSTM method due to it also influenced by several factors such as the size of the training data, the number of layers and neurons in the network, the type of activation function used, and the parameter tuning performed. Therefore, it is important to conduct a thorough evaluation and determine optimal parameters for each case of Bi-LSTM. The proposed method is applied to various optimizers and various numbers of hidden layers. By tuning hyper-parameter is addressed to improve classifier model for prediction [28]. Then, the lowest error rate is achieved at Adam optimizer with 32 hidden layers as shown in Table V.

Cross entropy is a loss function that serves to measure performance of the classification model. The cross Entropy will increase if probability predictions increasingly deviate from the actual class. The lower the cross-value entropy, the better is the model [29]. The value of the loss function appears to be homogeneous in deep learning algorithms including Bi-LSTM and LSTM as shown in the Fig. 11, Fig. 12, and Fig. 13.

Furthermore, the computational time of deep learning method is higher than the classical machine learning method due to complexity in classifier model as shown in Fig. 14. The Bi-LSTM algorithm is the highest computational time.

TABLE V. ERROR RATE OF TUNING HYPER-PARAMETER

LSTM#	Num of h_layers	*Adam opt*	*RMSprop*	*SGD*
1	8	0.0092005	0.0077283	0.0397047
1	16	0.0103672	0.0051495	0.0341338
1	32	0.0015641	0.000913	0.0276439
2	8	0.0085866	0.0057495	0.0699882
2	16	0.0034778	0.0007774	0.0567839
2	32	3.4505E-05	4.868E-05	0.0569696
3	8	0.0047429	0.0019380	0.0562194
3	16	0.0014209	0.0026316	0.0557157
3	32	7.5638E-05	0.0001729	0.0476583

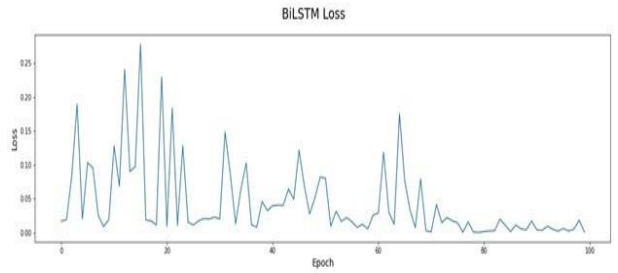

Fig. 11. The Bi-LSTM loss function.

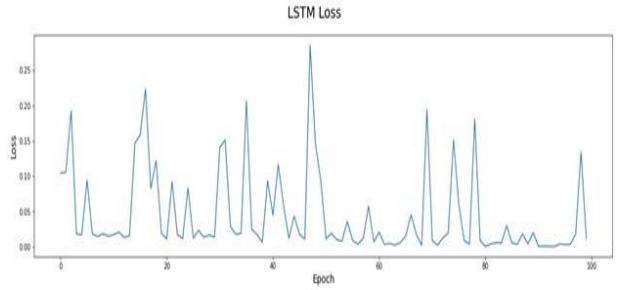

Fig. 12. The LSTM loss function.

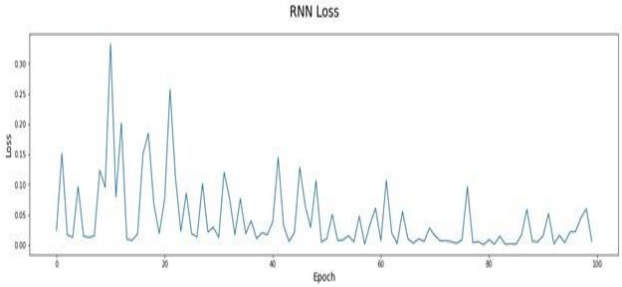

Fig. 13. The RNN loss function.

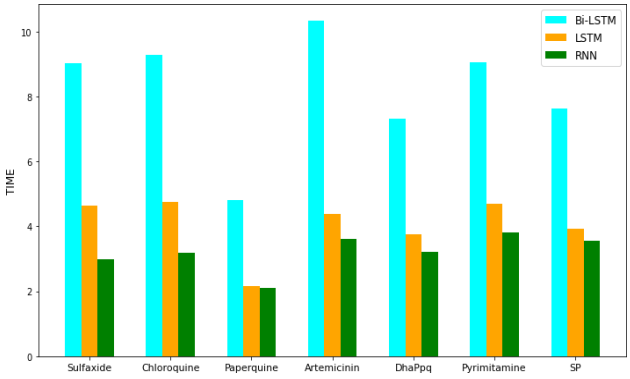

Fig. 14. Computational time comparison.

V. Conclusion

Predicting anti-malaria drug resistant based on DNA-Barcode sequences using Bi-LSTM method achieved high performance, especially for recall, and f1-score. However, the Bi-LSTM classification model requires a high computational time. Another lack of the proposed method is dependence on setting hyper-parameter of the method. Therefore, it is necessary to develop the tuning parameter to reduce the epoch of computational time for learning during the classifier model using training data.

Acknowledgment

This work was financially supported in part by the Faculty of Computer Science, Brawijaya University through a Grant of Doctoral Associate Professor "HDLK 2022" under contract number 2126/UN10.F15/PN/2022. Therefore, we are grateful for this funding and support of this research.

References

[1] S. Patgiri, K. Sarma, N. Sarmah, N. Bhattacharyya, D. Sarma, T. Nirmolia, D. Bhattacharyya, P. Mohapatra, D. Bansal, P. Bharti, et al., "Characteri- zation of drug resistance and genetic diversity of plasmodium falciparum parasites from Tripura, northeast India," Scientific Reports, vol. 9, no. 1, pp. 1–10, 2019.

[2] S. Dahlström, P. E. Ferreira, M. I. Veiga, N. Sedighi, L. Wiklund, A. Mårtensson, A. Färnert, C. Sisowath, L. Osório, H. Darban, et al., "Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in africa," The Journal of infectious diseases, vol. 200, no. 9, pp. 1456–1464, 2009.

[3] J. P. Gil and S. Krishna, "pfmdr1 (plasmodium falciparum multidrug drug resistance gene 1): a pivotal factor in malaria resistance to artemisinin combination therapies," Expert review of anti-infective therapy, vol. 15, no. 6, pp. 527–543, 2017.

[4] W. Deelder, E. D. Benavente, J. Phelan, E. Manko, S. Campino, L. Palla, and T. G. Clark, "Using deep learning to identify recent positive selection in malaria parasite sequence data," Malaria journal, vol. 20, no. 1, pp. 1–9, 2021.

[5] W. L. Hamilton, R. Amato, R. W. van der Pluijm, C. G. Jacob, H. H. Quang, N. T. Thuy-Nhien, T. T. Hien, B. Hongvanthong, K. Chindavongsa, M. Mayxay, et al., "Evolution and expansion of multidrug-resistant malaria in southeast asia: a genomic epidemiology study," The Lancet Infectious Diseases, vol. 19, no. 9, pp. 943–951, 2019.

[6] B. J. Neves, R. C. Braga, V. M. Alves, M. N. Lima, G. C. Cassiano, E. N. Muratov, F. T. Costa, and C. H. Andrade, "Deep learning-driven re-search for drug discovery: Tackling malaria," PLoS computational biology, vol. 16, no. 2, p. e1007025, 2020.

[7] H. Zhang, J. Guo, H. Li, and Y. Guan, "Machine learning for artemisinin resistance in malaria treatment across in vivo-in vitro platforms," Iscience, vol. 25, no. 3, p. 103910, 2022.

[8] G. W. Ashdown, M. Dimon, M. Fan, F. Sánchez-Román Terán, K. Witmer, D. C. Gaboriau, Z. Armstrong, D. M. Ando, and J. Baum, "A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens," Science advances, vol. 6, no. 39, p. eaba9338, 2020.

[9] A. Djimdé, O. K. Doumbo, J. F. Cortese, K. Kayentao, S. Doumbo,Y. Diourté, D. Coulibaly, A. Dicko, X.-z. Su, T. Nomura, et al., "A molec-ular marker for chloroquine-resistant falciparum malaria," New England journal of medicine, vol. 344, no. 4, pp. 257–263, 2001.

[10] L. Li, S. Yan, B. Lin, Q. Shi, and Y. Lu, "Single-cell proteomics for cancer immunotherapy," Advances in cancer research, vol. 139, pp. 185–207, 2018.

[11] S. A. Kazane, D. Sok, E. H. Cho, M. L. Uson, P. Kuhn, P. G. Schultz, and V. V. Smider, "Site-specific dna-antibody conjugates for specific and sensitive immuno-pcr," Proceedings of the National Academy of Sciences, vol. 109, no. 10, pp. 3731–3736, 2012.

[12] A. V. Ullal, V. Peterson, S. S. Agasti, S. Tuang, D. Juric, C. M. Castro, and R. Weissleder, "Cancer cell profiling by barcoding allows multiplexed protein analysis in fine-needle aspirates," Science translational medicine, vol. 6, no. 219, pp. 219ra9–219ra9, 2014.

[13] R. Weissleder and H. Lee, "Automated molecular-image cytometry and analysis in modern oncology," Nature Reviews Materials, vol. 5, no. 6, pp. 409–422, 2020.

[14] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, 2015.

[15] S. Khan, "Ethem alpaydin. Introduction to machine learning (adaptive computation and machine learning series). The Mit Press, 2004. isbn: 0

262 01211 1," Natural Language Engineering, vol. 14, no. 1, pp. 133–137, 2008.

[16] E. Kavlakoglu, "Ai vs. machine learning vs. deep learning vs. neural networks: What's the difference?" IBM, May, 2020.

[17] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional lstm recurrent neural network for forecasting network-wide traffic state with missing values," Transportation Research Part C: Emerging Technologies, vol. 118, p. 102674, 2020.

[18] K. Arunachalam, S. Thangamuthu, V. Shanmugam, M. Raju, and K. Prem- raj, "Deep learning and optimization for quality-of-service modelling," Journal of King Saud University-Computer and Information Sciences, 2022.

[19] T. Peng, C. Zhang, J. Zhou, and M. S. Nazir, "An integrated framework of bi-directional long-short term memory (bilstm) based on sine cosine algo- rithm for hourly solar radiation forecasting," Energy, vol. 221, p. 119887, 2021.

[20] A. Kulshrestha, V. Krishnaswamy, and M. Sharma, "Bayesian bilstm approach for tourism demand forecasting," Annals of tourism research, vol. 83, p. 102925, 2020.

[21] M. Cui, "District heating load prediction algorithm based on bidirectional long short-term memory network model," Energy, p. 124283, 2022.

[22] H. Jindal, N. Sardana, and R. Mehta, "Analyzing performance of deep learning techniques for web navigation prediction," Procedia Computer Science, vol. 167, pp. 1739–1748, 2020.

[23] D. P. Kingma, "&ba j.(2014). adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2015.

[24] C. V. Plowe, "Malaria chemoprevention and drug resistance: a review of the literature and policy implications," Malaria Journal, vol. 21, no. 1, p. 104, 2022.

[25] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," Annals of internal medicine, vol. 110, no. 11, pp. 916–921, 1989.

[26] J. Liang, "Confusion matrix," POGIL Activity Clearinghouse, vol. 3, no. 4, 2022.

[27] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sam- pling approach for imbalanced learning," in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pp. 1322–1328, IEEE, 2008.

[28] K. Shankar, Y. Zhang, Y. Liu, L. Wu, and C.-H. Chen, "Hyperparameter tuning deep learning for diabetic retinopathy fundus image classification," IEEE Access, vol. 8, pp. 118164–118173, 2020.

[29] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," Entropy, vol. 20, no. 3, p. 208, 2018.