ABSTRACT

Malaria is a disease that is dangerous to humans. Various attempts to find a drug that can overcome this disease. However, resistance from parasites to drugs is still found. Therefore we need an alternative medicine that can overcome this resistance problem. To find these alternative drugs, testing activities are needed in the laboratory. However, this requires a lot of time and money. One such alternative substance is fusidic acid which is known to have good potential to become a compound of anti-malaria agents. However, the IC50 value of fusidic acid compounds is still high so the optimization of fusidic acid derivatives is needed. In studies like this, the Quantitative Structure Activity Relationship (QSAR) method is commonly used. Mainly to predict the chemical properties of a compound. In hopes of creating predictive models of hereditary compounds that have better anti-malaria activity, the authors use the QSAR Method by Using Genetic Algorithms for feature selection and Artificial Neural Networks to create predictive models.

Keywords: Fusidic Acid, Quantitative Structure Activity Relationship, Genetic Algorithms, Artificial Neural Networks Introduction

ABSTRAKSI

Malaria adalah penyakit yang berbahaya bagi manusia. Berbagai upaya untuk menemukan obat yang dapat mengatasi penyakit ini. Namun, resistensi dari parasit terhadap obat masih ditemukan. Oleh karena itu diperlukan obat alternatif yang dapat mengatasi masalah resistensi ini. Untuk menemukan obat alternatif ini, kegiatan pengujian diperlukan di laboratorium. Namun, ini membutuhkan banyak waktu dan uang. Salah satu zat alternatif tersebut adalah asam fusidat yang dikenal memiliki potensi yang baik untuk menjadi senyawa agen anti-malaria. Namun, nilai IC50 dari senyawa asam fusidic masih tinggi sehingga optimalisasi turunan asam fusidic diperlukan. Dalam studi seperti ini, metode Quantitative Structure Activity Relationship (QSAR) umumnya digunakan. Terutama untuk memprediksi sifat kimia suatu senyawa. Dengan harapan menciptakan model prediktif senyawa herediter yang memiliki aktivitas anti-malaria yang lebih baik, penulis menggunakan Metode QSAR dengan Menggunakan Algoritma Genetika untuk pemilihan fitur dan Jaringan Syaraf Tiruan untuk membuat model prediksi.

Kata kunci : Asam Fusidic, Hubungan Aktivitas Struktur Kuantitatif, Algoritma Genetika, Pengantar Jaringan Syaraf Tiruan

1. INTRODUCTION

Nowadays there are still many adverse effects due to malaria outbreaks found in several regions. According to the WHO report, the impact of malaria has increased the socioeconomic burden substantially in some developing countries [2]. In 2013, there were 198 million cases of malaria which caused 58,000 deaths worldwide. The heaviest impact is experienced by residents in sub-Saharan Africa, especially children under the age of 5 years [2].

So far, malaria outbreaks can be overcome by using available drug regimens, such as artemisinin [3]. However, the regimen still has some shortcomings, mostly in terms of resistance to parasites. This is based on findings. Parasitic resistance to this regimen in several regions such as Cambodia, Vietnam, Myanmar, and Indonesia [4] - [6]. Therefore, the development of anti-malaria substances is needed, especially substances that do not have resistance to other drugs. Currently there are several potential substances for anti-malaria use, one of which is fusidic acid.

Fusidic acid is a type of compound commonly used in the medical world since 1960 [2]. This compound can control topical and systemic infections caused by Gram-positive bacteria such as Staphylococcus aureus [2], [3], [6], [7]. Several studies have proven that fusidic acid can act as a barrier to plasmodial EF-Gs located in the apicoplast and mitochondria [8]. However, it is necessary to ensure that the potential of these compounds is in the right value to inhibit malaria activity. IC50 is a measure of the potential of a substance in inhibiting certain biological or biochemical functions. IC50 is a quantitative measure that affects how much certain inhibitors (such as drugs) are needed to inhibit, in vitro, certain biological processes [9]. According to previous research, the IC50 value of fusidic acid is still too high to be used as an antimalarial [10]. Therefore a structural optimization process is needed to produce new derivatives with the anti-malaria activity that is better than fusidic acid.

In this regard, laboratory testing of these combined activities is required. However, evaluation of drug activity must have shortcomings in terms of time and cost. Therefore we need an alternative method that can be predicted before the activity is submitted in the laboratory. One method that can be used is the Quantitative Structure of Relationship Activity method or commonly referred to as QSAR. With the predictive ability of the QSAR method, the anti-malaria activity of a drug can be determined using molecular descriptors as input attributes.

2. MATERIALS

2.1 Related Works

The QSAR method has been widely used to predict the activity of anti-malaria agents. In 2000, Tonmunphean and colleagues conducted a QSAR study of 30 artemisinin-derived compounds with a linear regression method. Based on the results of their research, they succeeded in predicting one of the artemisinin properties accurately where a regression constant of -0.93 was obtained [11]. Then in 2006, Zahouily and colleagues used multiple linear regression and artificial neural network methods to predict the anti-malaria activity of 2-aziridinyl and 2,3-bis (aziridinyl) -1,4-naphthoquinonyl sulfonate and acylate derivatives. They concluded that the anti-malarial activity of the compound was highly dependent on the hydrophobic character, the recipient of hydrogen bonds, and the steric factor of the substituent. [12].

In 2013, Worachartcheewan and colleagues used the MLR and AJST methods to predict bisino-benzimidazole derivative activity. They determined that a model could be validated using a set of external coefficients (RExt = 0.9978 and 0.9844) and root meant an external square error set (RMSEExt = 0.0764 and 0.1302) as well as predictions from an external set (QExt2 = 0.9956 and 0.9690) which were good [13] . In 2015, Iman and colleagues conducted a study using the MLR method to determine molecular descriptors for the evaluation of predictions on the composition of Quinolone derivatives. From the results of his research, it can be concluded that the molecular descriptors of PJI2, Mv, PCR, nBM, and Var have a significant influence on anti-malaria activity in these compositions [14]. In 2014 Sharma and colleagues put together the annealing method and closest neighbors to predict the anti-malaria activity of azaaurone derivatives. They get pretty good prediction results with the parameter value q2 = 0.6906 and pred_r2 value = 0.7454 [15].

In 2014 Sahu conducted a QSAR study using genetic algorithm methods and multiple linear regression of the anti-malaria activity of a functioning 7-chloro-4-aminoquinolin bond. Based on the results of his research, he obtained an accurate prediction model with validation values of r2 = 0.9188, q2 = 0.8349, and pred_r2 = 0.7258. Subsequently, in 2015, Kalani used the MLR method to construct models and in vitro methods for validation of the antiplasmodial activity of ursolic acid derivatives. He concluded from the contraction of ursolic acid derivatives, UA-18 and UA-21 showed a significant dose-dependent value on antiplasmodial activity [16]. In addition, several other studies also use the QSAR method to predict the activity of anti-malarial agents [17] - [19].

In this study, we aim to build a QSAR model to predict fusidic acid derivatives as anti-malaria agents. A series of these compounds with activity values (IC50) were used to construct the model. these compounds are classified into active and inactive by using a certain limit value to the value of pIC50. The value of IC50 is converted to pIC50 as follow, pIC50 = -log(IC50). We calculate a number of 2-dimensional (2D) molecular descriptors of fusidic acid derivatives to give the numerical attributes of the inhibitors. The descriptors used were selected using the genetic algorithm (GA) method to produce a set of descriptors with low cross-entropy loss values. Here we use the GA method because this method is usually used in the QSAR Analysis feature selection procedure. [20, 21, 22] The GA method provides a set of features that produce more accurate results compared to other methods, such as forward selection or backward elimination. [23] Finally, the QSAR model was built to connect descriptors with a combined class using neural networks (NN) with optimized hyperparameters.

2.2 Genetic Algorithm

Genetic algorithm (AG) was introduced by Holland in the early 1970s as an optimization approach by modeling the simulation of the evolutionary process of living species [25]. This method follows Darwin's classic rules of natural evolution and uses random methods to obtain optimal nonrandom solutions [26].

2.3 Artificial Neural Network

Artificial Neural Networks (ANN) are mathematical models that try to simulate the structure and function of biological neural networks. The basic building blocks of each artificial network are artificial neurons, which are simple mathematical models (functions). Such a model has three simple sets of rules: multiplication, addition, and activation. At the entrance, the input of artificial neurons is weighed as to what if each input value is multiplied by individual weights. In the middle, artificial neurons that present a number of functions that add up all the inputs and have been weighted. At the exit, the neuron is made to represent the amount of input and which has previously been weighted and will pass the activation function which is also called the transfer function.

2.4 Model Validation

We assess the performance of our model by evaluating the accuracy of prediction of binary classification with several statistical parameters, e.g. true positive (TP), false positive (FP), true negative (TN), false negative (FN), sensitivity (SE), specificity (SP), overall predictive accuracy (Q), and Matthews correlation coefficient (MCC). Here, sensitivity measures the ability of the model to recognize data when positive conditions are present, while specificity measures the ability of the model to correctly exclude data when conditions do not exist. MCC is a measure of the quality of binary classification. MCC is a coefficient that represents the correlation between the observed and predicted binary classifications. So, MCC can be used to measure the overall quality of binary classification. The parameter values are evaluated using Equations 2.

$$
\begin{aligned}
\mathrm{SE} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
\mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\
\mathrm{Q} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+FN} \\
\mathrm{MCC} &= \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}}
\end{aligned}
$$

Equations 2.

3. METHODS

3.1 Feature Selection

A total of 3488 2-dimensional (2D) descriptors will be selected using several methods to obtain the most significant features. The molecular descriptors represent the structure, topology, and electrostatic properties of each inhibitor. The procedure for selecting descriptors is done in two cups. For the first time, a statistical analysis is performed to select the chosen descriptors. Descriptors with zero variation values or containing zero values of more than 25% are required. Then, an analysis of fixing Pearson is carried out to reduce bias and improve descriptors with similar information. Descriptors that have weaknesses (Pearson conversion coefficient <0.1) with pIC50 or have strong strengths (contrary to Pearson coefficients> 0.4) with other descriptors are moved. Two among the descriptors who have strong pIC50 agreements will be chosen based on their correlation with pIC50.

Furthermore, the set of descriptors obtained is standardized to obtain descriptor values in vulnerable 0 to 1. The standardization procedure is not important to ensure all descriptors are within the same average range and standard deviation. This condition is necessary because molecular descriptors will be used in the neural network (ANN) methods.

In the second step, the best combination of descriptors is chosen using the Genetic Algorithm (GA) method. The selection of descriptors begins with the approved chromosome as a collection of bits. The total amount is equal to the number of descriptors to choose from. The descriptor bit value will be '1' if the descriptor is selected, if not '0'. The author repeats the AG procedure with 100 iterations and stores the 20 best chromosomes in each iteration. To produce a new chromosome, the writer performs a crossover procedure at each bit point of the chromosome. The resulting chromosome will be accepted if it contains the desired description and is different from the chromosomes in the participation. Then, the mutation process is carried out with a probability of 0.2.

3.2 Optimization and Prediction Model

Based on the principle of work and the separate rules of artificial neurons it looks like nothing special, the full potential and power of the calculation of this model will be active compilation we start connecting it to artificial neural networks (Figure 1) This artificial neural network uses basic facts and simple rules.

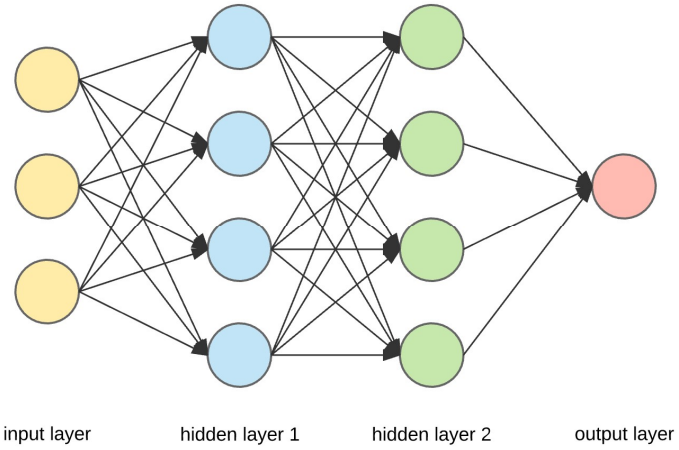

Figure 1: The working principle of artificial neurons

To improve the performance of ANN, the authors make hyperparameters consisting of network structure parameters, such as hidden layers, hidden units, dropout rates, and hyperparameter agitational training, such as learning levels and the number of layers. The details of hyperparameter candidates will shown at Table 1.

Table 1 : Hyperparameter Range

Hyperparameter name	Dropout rate
Neurons	[5, 6, 7, 8, 9, 10]
Learning Rate	[0.001, 0.01, 0.1]
Dropout Rate	[0.0, 0.1, 0.2]
Momentum	[0.0, 0.1, 0.2]

3.3 Validation Method

The models are also evaluated by visualizing the receiver operating characteristic (ROC) curve. This curve illustrates the success rate and error prediction observed in the classification model. The ROC curve is plotted by taking true positive levels and false-positive levels on the y-axis and x-axis, each. The accuracy of predictions can be easily recognized by analyzing the characteristics of the curve. [25] From this curve, we can also calculate other parameters, namely the area below ROC curve (AUC). AUC measures the model's ability to distinguish between the two classification groups and thus represents the accuracy of predictions.

After we get the model with the best performance, the scramble experiment is performed on the best model to prove that the performance of the model is not appropriate for coincidence

correlation. 10 random models are generated by shuffling target value while preserving descriptors. Scrambled model performance is evaluated by calculating MCC values and comparing values with non-randomized models.

4. RESULT AND DISCUSSION

4.1 Dataset Diversity

After determining the IC50 mean value, a number is obtained as an indicator of class separation. the target will be given a value of 1 if the composition ic50 exceeds the indicator value. and also given a value of 0 if the composition is less than the indicator value. after that, all compounds will be divided into two types of data sets randomly with a test size of 0.2.

4.2 Desriptor Selection

We chose descriptors using statistical analysis methods and genetic algorithms (GA). The number of descriptors decreased from 3340 to 1178 after we removed descriptors with zero variance and descriptors that had low standard deviation values. Then, by removing descriptors with weak correlations with targets (pIC50) and strong correlations with other descriptors, we obtained 51 descriptors. We assume that the chosen descriptor contains enough information to be used in the model.

We obtained 5 models from GA selection which are summarized in Table 2. And also, we obtained log loss information which are summarized in figure 1. Some descriptors, such as ATSC1s, nHBint7, TDB5i, TDB8s, RDF45m, and RDF85m were found selected in several models. This shows that the descriptors have valuable information that can be used to build the model.

Figure 1 : The plot of log loss as a function of iteration

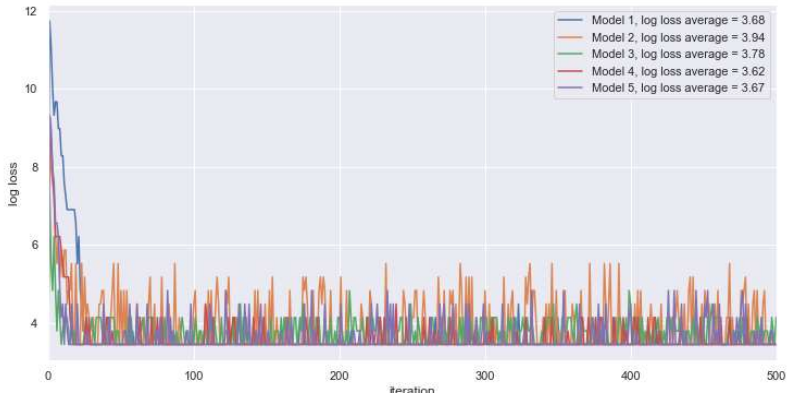

Table 2 : The summary of the model obtained from GA selection.

Model No.	Number of desc.	Selected desc.	Loss
1	5	ATSC1dv, **SMR_VSA1**, **nAtomLC**, PPSA-3, **RDF115m**	3.69
2	6	**SMR_VSA1**, EState_VSA7, **VE3_Dzp.1**, SHBint8, **PNSA-3**, **RDF95m**	3.95
3	7	ATSC2s, ATSC7are, SdO, **PNSA-3**, **RDF95m**, **RDF115m**, **RDF45s**	3.79
4	8	**PEOE_VSA1**, **PEOE_VSA10**, **SMR_VSA1**, **EState_VSA3**, **VE3_Dzp.1**, **nAtomLC**, **RDF115m**, **RDF45s**	**3.63**
5	9	**PEOE_VSA1**, **PEOE_VSA10**, **SMR_VSA1**, **EState_VSA3**, VE3_DzZ.1, **VE3_Dzp.1**, **nAtomLC**, **RDF115m**, **RDF45s**	3.68

4.3 Neural Network Model

We create neural network models for 5 descriptor combinations obtained from the genetic algorithm (GA) selection. Hyperparameter adjustment is performed to obtain optimal hyperparameter values and improve the performance of each model. The results of the hyperparameter adjustment process are presented in Table 3. We find that the number of optimized layer numbers for all models is 5. Because this number is the maximum number of layers, this shows that increasing the number of layers increases the performance of the NN model. Meanwhile, the optimized hidden unit values are found to be different for each model.

Table 3 : The summary of hyperparameter tuning of NN model.

Model no.	Dropout rate	Learn rate	Momentum	Neurons
1	0,2	0.010	0.1	9
2	0.1	0.100	0.2	10
3	0.1	0.001	0.1	7
4	0.2	0.010	0.0	6
5	0.0	0.100	0.1	10

4.4 Validation Result

The results of model performance validation are presented in Table 4. In the case of training data, we found that all models can accurately predict target values, which are indicated by high MMC values. By using model 4, the resulted values of SE, SP, Q, and MCC are 0.68, 0.70, 0.79 and 0.56, respectively. Even so, model 4 has an overfitting tendency because it has an MCC value on the test data that is close to zero.

The test set validation is found to produce the values of the validation parameter that are lower than training set validation. According to the results, we found that the test set prediction by using model 2 is more accurate than other models. this is indicated by MCC values on the train set and test set which are higher and more stable than other models.

Table 4 : The summary of validation results of the prediction on training and test set.

Model	TP	FP	TN	FN	SE	SP	PR	Q	F1- score	AUC	MCC
Training Set											
1	25	4	9	10	0.71	0.69	0.86	0.73	0.72	0.80	0.42
2	27	2	10	9	0.75	0.83	0.93	0.75	0.73	0.80	0.47
3	29	0	17	2	0.94	1.00	1.00	0.65	0.54	**0.85**	0.26
4	26	3	7	12	0.68	0.70	0.90	**0.79**	**0.79**	0.84	**0.56**
5	26	3	9	10	0.72	0.75	0.90	0.75	0.74	0.84	0.47
Test set											
1	3	4	1	5	0.38	0.20	0.43	0.62	0.60	0.71	0.28
2	4	3	1	5	0.44	0.25	0.57	**0.69**	**0.69**	0.74	**0.41**
3	6	1	6	0	1.00	0.86	0.86	0.46	0.34	0.55	-0.27
4	3	4	2	4	0.43	0.33	0.43	0.54	0.53	0.64	0.10
5	3	4	2	4	0.43	0.33	0.43	0.54	0.53	**0.76**	0.10

Then, the y-scrambling analysis is performed on model 2 to confirm that the results do not match the coincidence correlation. The MCC value of model 2 which is used to predict 10 random data sets is presented in Figure 2. For comparison reasons, we also present the MCC value of model 2 for data that is not randomized. We found that the MCC value for 10 randomized data was lower than the original MCC data. This confirms that the accuracy of model 2 is not related to chance correlation

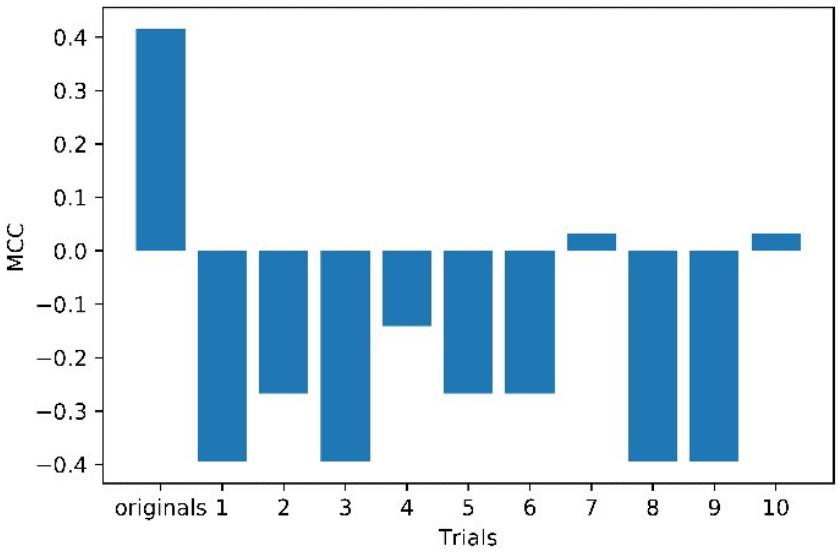

Figure 2 : Y-scrambling result of model 2.

5. CONCLUCION

We construct neural network models for 5 descriptor combinations obtained from the genetic algorithm (GA) selection. Hyperparameter adjustment is performed to obtain optimal hyperparameter values and improve the performance of each model. The results of the hyperparameter adjustment process are presented in Table 3. Because this number is the maximum number of layers, this shows that increasing the number of layers increases the performance of the NN model. Meanwhile, the optimized hidden unit values are found to be different for each model.

REFERENCES

[1] N. K. Sahu, M. C. Sharma, V. Mourya, and D. V. Kohli, "QSAR studies of some side chain modified 7-chloro-4-aminoquinolines as antimalarial agents," Arab. J. Chem., vol. 7, no. 5, pp. 701–707, Nov. 2014

[2] "World Malaria Report 2014 - Summary," p. 28, 2014.

[3] T. E. Wellems and C. V. Plowe, "Chloroquine-Resistant Malaria," p. 7.

[4] W. R. J. Taylor and N. J. White, "Antimalarial Drug Toxicity," Drug Saf., vol. 27, no. 1, pp. 25–61, Jan. 2004.

[5] E. A. Ashley et al., "Spread of Artemisinin Resistance in Plasmodium falciparum Malaria," N. Engl. J. Med., vol. 371, no. 5, pp. 411–423, Jul. 2014.

[6] E. Tjitra et al., "Multidrug-Resistant Plasmodium vivax Associated with Severe and Fatal Malaria: A Prospective Study in Papua, Indonesia," PLOS Med., vol. 5, no. 6, p. e128, Jun. 2008.

[7] A. M. Dondorp et al., "Artemisinin Resistance in Plasmodium falciparum Malaria," N. Engl. J. Med., vol. 361, no. 5, pp. 455–467, Jul. 2009.

[8] W. O. Godtfredsen, S. Jahnsen, H. Lorck, K. Roholt, and L. Tybring, "Fusidic Acid: a New Antibiotic.," Nature, vol. 193, 1962.

[9] Tom B, "Chapter 7 - Drug–Drug Interactions: Part One (Small Molecule Drugs)" Science Direct.

[10] G. Kaur et al., "Synthesis of fusidic acid bioisosteres as antiplasmodial agents and molecular docking studies in the binding site of elongation factor-G," MedChemComm, vol. 6, no. 11, pp. 2023–2028, 2015.

[11] S. Tonmunphean, V. Parasuk, and S. Kokpol, "QSAR Study of Antimalarial Activities and Artemisinin-Heme Binding Properties Obtained from Docking Calculations," Quant. Struct.-Act. Relatsh., vol. 19, no. 5, pp. 475–483, 2000.

[12] M. Zahouily, M. Lazar, A. Elmakssoudi, J. Rakik, S. Elaychi, and A. Rayadh, "QSAR for anti-malarial activity of 2-aziridinyl and 2,3-

[13] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, "QSAR study of amidino bis-benzimidazole derivatives as potent anti-malarial agents against Plasmodium falciparum," Chem. Pap., vol. 67, no. 11, pp. 1462–1473, Nov. 2013.

[14] M. Iman, A. Davood, and A. Khamesipour, "Computational Study of Quinolone Derivatives to Improve their Therapeutic Index as Anti-malaria Agents: QSAR and QSTR," Iran. J. Pharm. Res. IJPR, vol. 14, no. 3, pp. 775–784, 2015.

[15] M. C. Sharma, S. Sharma, P. Sharma, and A. Kumar, "Pharmacophore and QSAR modeling of some structurally diverse azaaurones derivatives as anti-malarial activity," Med. Chem. Res., vol. 23, no. 1, pp. 181–198, Jan. 2014.

[16] K. Kalani, H. Singh Cheema, H. Tripathi, F. Khan, M. P. Daroker, and S. Kumar Srivastava, "QSAR-guided semi-synthesis and in vitro validation of antiplasmodial activity in ursolic acid derivatives," RSC Adv., vol. 5, no. 41, pp. 32133–32143, 2015.

[17] V. H. Masand, D. T. Mahajan, P. Gramatica, and J. Barlow, "Tautomerism and multiple modelling enhance the efficacy of QSAR: antimalarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl," Med. Chem. Res., vol. 23, no. 11, pp. 4825–4835, Nov. 2014.

[18] T.-S. Tseng, Y.-C. Lee, N.-W. Hsiao, Y.-R. Liu, and K.-C. Tsai, "Comparative study between 3D-QSAR and Docking-Based Pharmacophore models for potent Plasomodium falciparum dihydroorotate dehydrogenase inhibitors," Bioorg. Med. Chem. Lett., vol. 26, no. 2, pp. 265–271, Jan. 2016.

[19] T. P. Jiménez Villalobos, R. Gaitán Ibarra, and J. J. Montalvo Acosta, "2D, 3D-QSAR and molecular docking of 4(1H)-quinolones analogues with antimalarial activities," J. Mol. Graph. Model., vol. 46, pp. 105–124, Nov. 2013.

[20] Hasegawa K, Miyashita Y and Funatsu K 1997 Journal of Chemical Information and Computer Sciences 37306–310 ISSN 0095-2338 URL https://doi.org/10.1021/ci960047x

[21] Sagrado S and Cronin M T D 2008 Analytica Chimica Acta 609 169–174 ISSN 0003-2670 URL http://www.sciencedirect.com/science/article/pii/S0003267008000846

[22] Gao H, Lajiness M S and Drie J V 2002 Journal of Molecular Graphics and Modelling 20 259–268 ISSN 1093-3263 URL http://www.sciencedirect.com/science/article/pii/S109332630100122X

[23] Xu L and Zhang W J 2001 Analytica Chimica Acta 446 475–481 ISSN 0003-2670 URL http://www.sciencedirect.com/science/article/pii/S0003267001012715

[24] E. Alizadeh, S. M. Lyons, J. M. Castle, and A. Prasad, “Measuring systematic changes in invasive cancer cell shape using Zernike moments,” Integr. Biol., vol. 8, no. 11, pp. 1183–1193, Nov. 2016

[25] S. Tonmunphean, V. Parasuk, and S. Kokpol, “QSAR Study of Antimalarial Activities and Artemisinin-Heme Binding Properties Obtained from Docking Calculations,” Quant. Struct.-Act. Relatsh., vol. 19, no. 5, pp. 475–483, 2000.

[26] D. E. Goldberg and J. H. Holland, “Genetic Algorithms and Machine Learning,” Mach. Learn., vol. 3, no. 2, pp. 95–99, Oct. 1988.