A Robust SMILES-Based Prediction of Aqueous Solubility of Diverse Antiplasmodial Compounds using Machine Learning Algorithms

Thecla O. Ayoka 1* , Charles O. Nnadi 2,*

1Department of Science Laboratory Technology (Biochemistry Unit), Faculty of Physical Sciences, University of Nigeria, Nsukka, 410001, Enugu State, Nigeria; thecla.ayoka@unn.edu.ng

2Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, 410001, Enugu State, Nigeria; charles.nnadi@unn.edu.ng

* Correspondence: thecla.ayoka@unn.edu.ng (T.O.A.); charles.nnadi@unn.edu.ng (C.O.N.)

Abstract

Apart from the pharmacodynamics of drugs and the resistance of the *Plasmodium falciparum* parasite to existing antimalarial drugs, pharmacokinetic-related properties of drugs also hamper their translation. The need to develop novel drugs with optimum solubility profiles necessitated the training of an efficient machine learning regression model for the prediction of the solubility of a series of compounds. Four descriptors: octanol-water partition coefficient, molecular weight, number of rotatable bonds and aromatic proportion from the simplified molecular-input line-entry system (SMILES) of 11,478 antiplasmodial molecules were used. This was trained using five regression models; multiple linear regression, k-nearest neighbors, LASSO regression, support vector regressor and random forest regressor (RFR)) to predict the solubility of molecules. The evaluation metrics (R^2, mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE)) were used to assess the model performance. Of the performed algorithms, the RFR produced a robust model with model statistics of MSE 0.54, R^2 0.85, MAE 0.41 and RMSE 0.73. The *F*-statistic for the model was 7214, showing a strong correlation between the descriptors and solubility of molecules. This could efficiently predict the antimalarial activity for untested molecules to select promising ligands as leads for further optimization.

Keywords: Antimalarial; machine learning; molecule descriptors; regression models; solubility

Introduction

The WHO estimates that 50% of all malaria deaths among children under five occur in Africa [1]. Malaria is one of the most severe and dangerous diseases due to its detrimental and harmful effects on world populations, despite the organization's continued efforts to make treatments more available. The success of malaria control, prophylaxis, and treatment depends on the effectiveness of first-line artemisinin-based combination therapy (ACT), which is constantly threatened by the emergence and spread of drug resistance [2].

Due to the difficulties that arise during the process, the development of a drug typically takes 14 years from the requisite pre-clinical testing to regulatory approval [3]. Surprisingly, the most expensive component of drug development is conducting clinical trials. Given these challenges, predictive modeling techniques are projected to be more crucial in order to identify risks and establish strategies for the development of new anti-malarial drugs.

SMILES-based models for the prediction of aqueous solubility of various antiplasmodial chemical entities were created using machine learning techniques to prevent failures in the later stages of drug development [4].

This study has demonstrated that even with a sparse use of the descriptors, it is still possible to produce resilient, trustworthy, and accurate models. In the study, descriptor values for the compounds were extracted, analyzed to determine which descriptors are more significant for solubility, and then trained the models using Machine Learning techniques to determine which four implemented models had comparable performance

Materials and Methods

- *Dataset*

The antimalarial molecule list was obtained from literature and the public database ChEMBL. They were converted into their respective and appropriate SMILES using the PubChem Identifier Exchange Service. It consists of 11481 molecules, and three (3) features. After filtering out missing and duplicate values, a total of 11478 antimalarial molecules were left.

- *Data pretreatment*

The SMILES string was converted to rdkit object using the RDKit library. Thereafter, four molecular descriptors- cLogP (octanol-water partition coefficient), MW (Molecular weight), nRB (Number of rotatable bonds), and ArP (Aromatic proportion = the number of aromatic atoms/numbers of heavy atoms) were calculated from SMILES [5]. The computed descriptors were combined to form the X-matrix, while the LogS column of the dataset forms the Y-matrix.

- *Data split*

The X- and Y-matrix were split into train set (9182 molecules) and test set (2296 molecules), using a split ratio of 0.2, where 80 % is assigned to the train set and 20 % is assigned to the test set. The size of the training dataset was denoted as X-train, Y-train, while the size of the test dataset was X-test, Y-test. The training set was used to train the model, while 2296 molecules belonging to the test set were used to validate the models. The models were trained on the training set using the fit method [6].

- *Building regression models*

Five (5) machine learning scikit-learn algorithms (Multiple Linear Regression (MLR), *k*-Nearest Neighbours (*k*NN), LASSO regression, Support Vector Regressor (SVR) and Random Forest Regressor (RFR)) were deployed to predict the solubility of molecules, to select further viable lead ligands for creative drug design. These learning algorithms combine ensemble learning techniques and conventional learning techniques with linear and nonlinear methods [7]. The goal was to discover the best algorithm capable of predicting the antimalarial activity for untested compounds.

- *Model evaluation*

Different evaluation metrics (coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE) [7] were deployed to assess the performance of the models.

Results

Machine Learning Models

A multiple regression model was built to predict continuous solubility values of antimalarial compounds. Four x values were used and each feature was viewed as a dimension. Only four features were considered to predict the output using the multiple linear regression equation 1. This was derived from the coefficients of the features and the intercept.

$$LogS = 0.90 - 0.54LogP - 0.009MW + 0.038RB - 1.61AP \quad - \quad -1$$

Where LogS = Solubility of antimalarial compounds

LogP = Octanol-water partition coefficient

MW = Molecular weight

nRB = Number of rotatable bonds

ArP = Aromatic proportion

To prove further confidence in our predicted solubility values, the predicted solubility scores were plotted against the experimental solubility scores for both the train set and the test set, using different machine learning models (Figure 1). The closeness of the predicted solubility scores and the experimental scores shows the robustness of our ML models. This shows that the predictive powers of the models are competent.

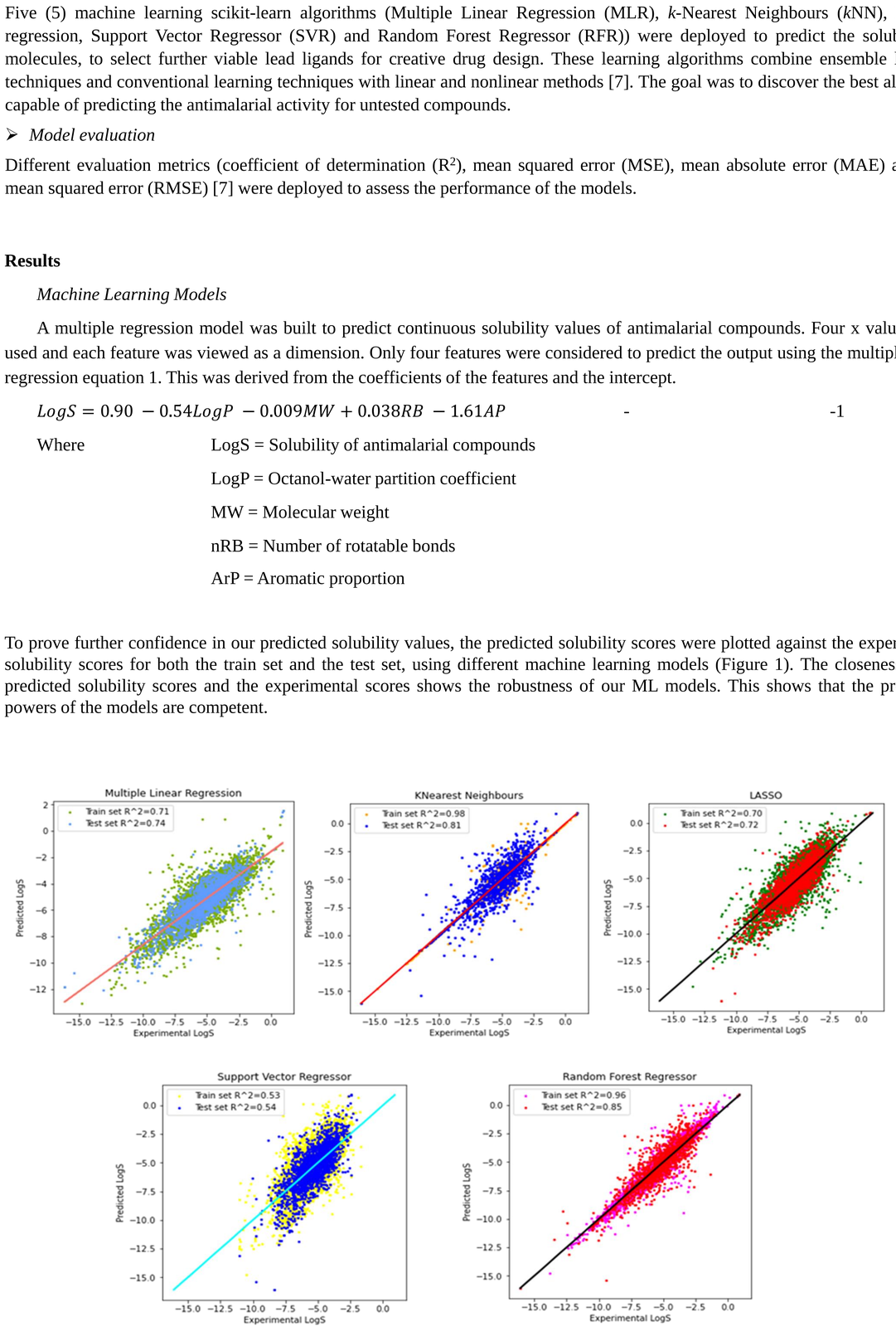

Figure 1. Correlation of the predicted and experimental solubility of antimalarial molecules in both the train and test data using different models.

Model evaluation and comparison

The summary of the performance of the models is shown in Figure 2. To find if the linear model is good, MSE and R^2 statistics are used. The lower the residue error, the better the model fits (the closer the data is to a linear relationship). R^2 measures the percentage of solubility variation that can be explained by descriptors.

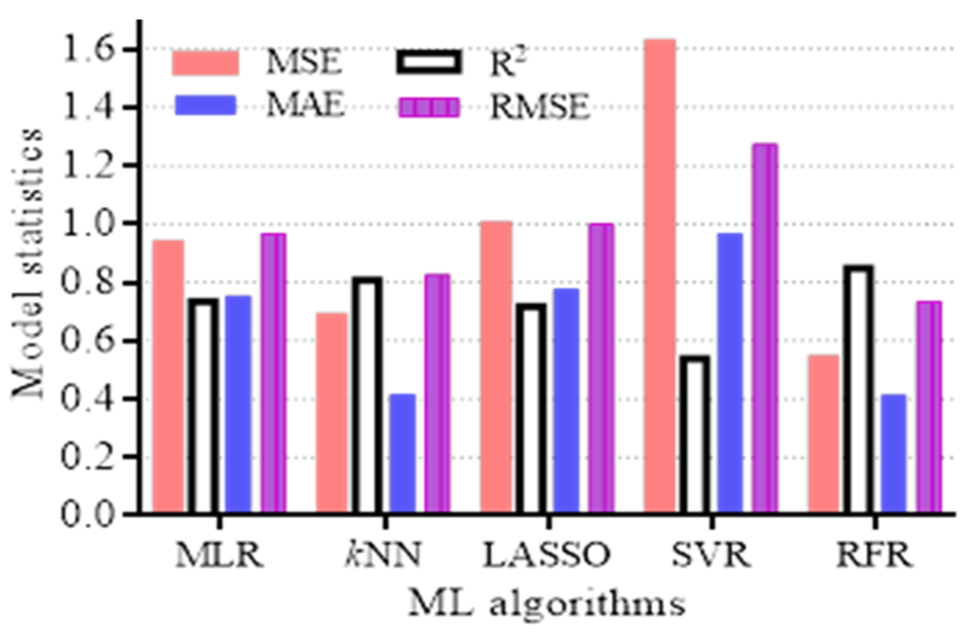

Figure 2. Grouped plots showing the models statistics values of the test set.. RFR = random forest regressor; SVR = support vector regressor; *k*-NN = KNearest neighbours; MLR = multiple linear regression

The size of train data is correlated to how well the learning algorithm function. Since the random forest regressor algorithm performed the best in the final test, its learning curves were examined.

Figure 3 displays the model performance and accuracy variations between the training dataset and the test dataset as the data volume varied. The learning curves' results revealed a consistent pattern in the four metrics. As the data volume increased, the model's performance on the test dataset gradually improved.

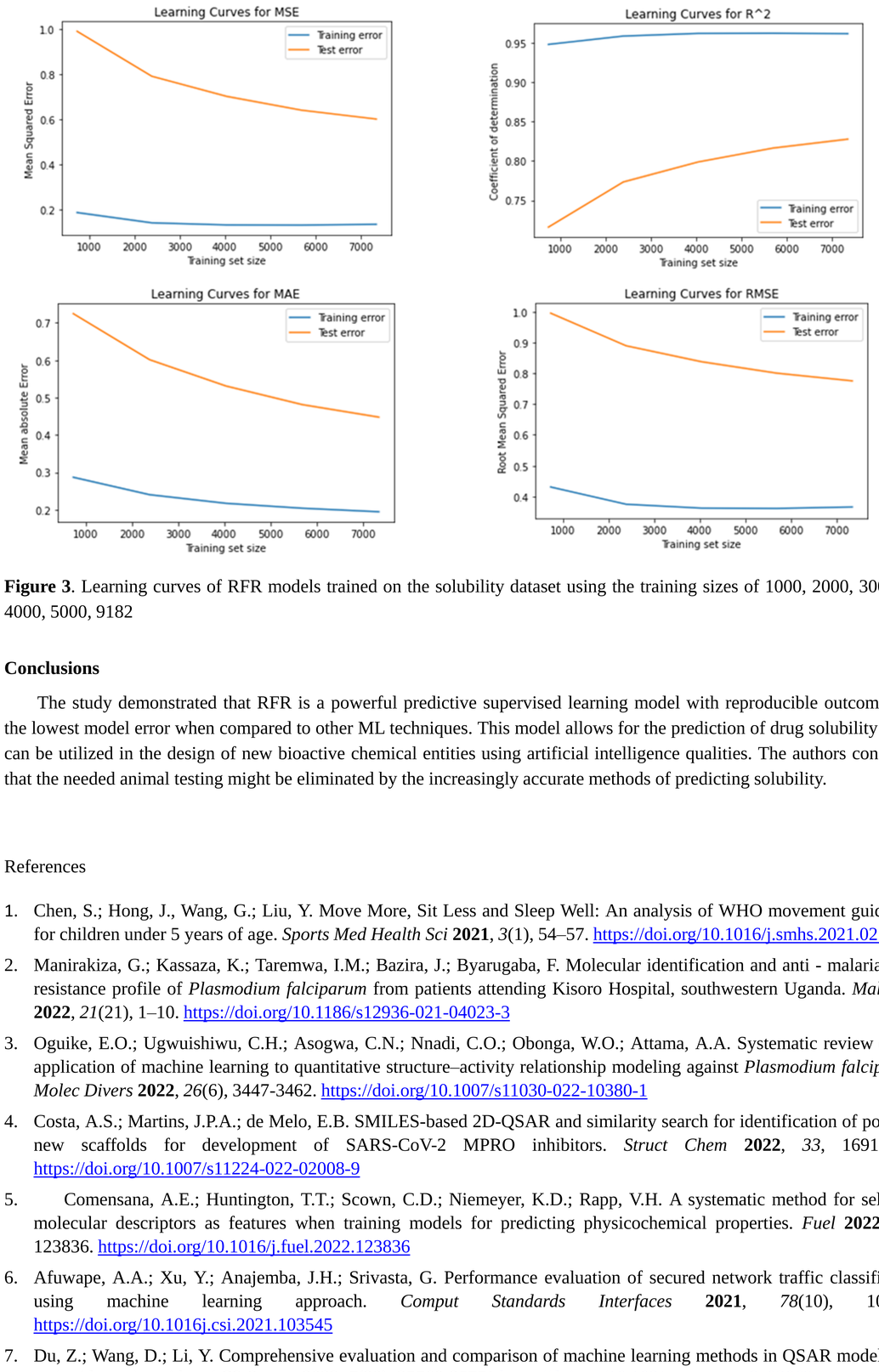

Figure 3. Learning curves of RFR models trained on the solubility dataset using the training sizes of 1000, 2000, 3000, 4000, 5000, 9182

Conclusions

The study demonstrated that RFR is a powerful predictive supervised learning model with reproducible outcomes and the lowest model error when compared to other ML techniques. This model allows for the prediction of drug solubility which can be utilized in the design of new bioactive chemical entities using artificial intelligence qualities. The authors concluded that the needed animal testing might be eliminated by the increasingly accurate methods of predicting solubility.

References

1. Chen, S.; Hong, J., Wang, G.; Liu, Y. Move More, Sit Less and Sleep Well: An analysis of WHO movement guidelines for children under 5 years of age. *Sports Med Health Sci* **2021**, *3*(1), 54–57. https://doi.org/10.1016/j.smhs.2021.02.003
2. Manirakiza, G.; Kassaza, K.; Taremwa, I.M.; Bazira, J.; Byarugaba, F. Molecular identification and anti - malarial drug resistance profile of *Plasmodium falciparum* from patients attending Kisoro Hospital, southwestern Uganda. *Malaria J* **2022**, *21*(21), 1–10. https://doi.org/10.1186/s12936-021-04023-3
3. Oguike, E.O.; Ugwuishiwu, C.H.; Asogwa, C.N.; Nnadi, C.O.; Obonga, W.O.; Attama, A.A. Systematic review on the application of machine learning to quantitative structure–activity relationship modeling against *Plasmodium falciparum*. *Molec Divers* **2022**, *26*(6), 3447-3462. https://doi.org/10.1007/s11030-022-10380-1
4. Costa, A.S.; Martins, J.P.A.; de Melo, E.B. SMILES-based 2D-QSAR and similarity search for identification of potential new scaffolds for development of SARS-CoV-2 MPRO inhibitors. *Struct Chem* **2022**, *33*, 1691-1706. https://doi.org/10.1007/s11224-022-02008-9
5. Comensana, A.E.; Huntington, T.T.; Scown, C.D.; Niemeyer, K.D.; Rapp, V.H. A systematic method for selecting molecular descriptors as features when training models for predicting physicochemical properties. *Fuel* **2022**, *321*, 123836. https://doi.org/10.1016/j.fuel.2022.123836
6. Afuwape, A.A.; Xu, Y.; Anajemba, J.H.; Srivasta, G. Performance evaluation of secured network traffic classification using machine learning approach. *Comput Standards Interfaces* **2021**, *78*(10), 103545. https://doi.org/10.1016j.csi.2021.103545
7. Du, Z.; Wang, D.; Li, Y. Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides. *ACS Omega* **2022**, *7*(29), 25760-25771 https://doi.org/10.1021/acsomega.2c03062