Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

A supervised classifier based chemoinformatics model to predict inhibitors essential for sexual reproduction and transmission of the *P. falciparum* parasite into mosquitoes

Asif Hassan Syed *, Tabrej Khan

Department of Computer Science, Faculty of Computing and Information Technology Rabigh (FCITR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 20 April 2019
Received in revised form
4 August 2019
Accepted 5 August 2019

Keywords:
Falciparum malaria
High-throughput screening dataset
Supervised learning based model
Precision
Recall

ABSTRACT

The *falciparum* malaria is a significant life-threatening disease caused by Plasmodium *falciparum* a protozoan parasite transmitted by the female Anopheles mosquito. The resistance of *P. falciparum* parasite to a limited class of antimalarial medicine has accelerated the process of screening a novel drug for *falciparum* malaria. In recent years the implementation of Machine Learning (ML) approaches to build a predictive model to facilitate the target-specific drug discovery process for both infectious and non-infectious pathogen has gained significance. The availability of High-throughput Screening (HTS) anti-malarial bioassay dataset has provided an opportunity to build ML-based chemoinformatics, predictive models, using features extracted from different Feature Selection (FS) algorithms. In the present study, a combination of feature selection algorithms namely Greedy Stepwise algorithm in association with CfsSubsetEval and Principal Components Analysis (PCA) in conjunction with Ranker method was used on the HTS dataset. The dataset comprising of *P. Falciparum* Calcium-Dependent Protein Kinase4 (PfCDPK4) inhibitors and non-inhibitors were used to train and build four state-of-art classifiers based model for predicting inhibitors of PfCDPK4 protein from an independent test dataset accurately. The classification models were evaluated based on specific statistical measures of the Weka software tool. The J48 classifier based predictive model was found to accurately predict active anti-PfCDPK4 molecule based on better Accuracy, Recall, Precision, and Area under the Curve (AUC) values. Thus, the authors conclude that the J48-based classification model will be efficient and cost-effective in screening future active anti-CDPK4 molecule against *P. falciparum* malaria parasite.

1. Introduction

P. falciparum a protozoan parasite is the causative agent of *falciparum* malaria and is transmitted to human by the female Anopheles mosquitos. As per the World Health Organization (WHO) malarial report of 2017, there were an assessed 216 million reported cases of malaria in ninety-one countries and approximately 445000 death cases reported globally until November 2017. The African region with 90 % malarial cases and 91 % death due to disease shares a majority malarial burden of the world (WHO, 2017). In recent times, the evolution of multi-drug resistant strains of *P. falciparum* against conventional antimalarial drugs namely chloroquine and primaquine (Trenholme and Carson, 1978; Mehta and Das, 2006) have resulted in limited therapeutic option for the treatment of *falciparum* malaria (Wongsrichanalai et al., 1992; Wongsrichanalai et al., 2002; Dua et al., 2003; Yang et al., 2011). With only eight medications in preclinical trials and only 13 new antimalarial drugs under clinical trial, the elimination of the multidrug-resistant strain of *P. falciparum* will not be easy. Therefore, for designing new next-generation antimalarial novel medicines for the removal of recently evolved *falciparum* malarial parasite will require target-based rapid screening of novel antimalarial hit molecules (Burrows et al., 2013). One of the significant hurdles in the detection of novel antimalarial molecules is the considerable cost

* Corresponding Author.
Email Address: shassan1@kau.edu.sa (A. H. Syed)
https://doi.org/10.21833/ijaas.2019.10.011
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7288-3098

associated with selection of novel hit molecules via in vitro and in vivo procedures.

In this regard, the availability of large dataset of the chemical molecule with their chemical structure and clear bio-activity information available via automated high-throughput bioassay (Wang et al., 2009) has helped the development of ML-based computational models to predict the bio-activity of a synthetic chemical molecule. The computational predictive model predicts the bio-activity of a chemical molecule based on the correlation between the chemical properties and the activity of the molecule (Schierz, 2009; Melville et al., 2009). The ML-based *in silico* virtual screening protocol enables rapid screening of target-based inhibitors from vast chemical molecule library. Therefore, the ML-based computational model enhances the rate of detection of active hit molecules thereby reducing the high-cost involved in the discovery of hit molecules from conventional automated high-throughput bioassay drug screening protocols. In the recent past many research groups have developed many ML-based predictive models to (1) predict anti-malarial molecules that inhibit apicoplast formation in *Plasmodium falciparum* (Jamal et al., 2013; Dixit and Singla, 2017; Bharti and Lynn, 2017), (2) screen anti-malarial hit molecules against Aspartyl aminopeptidase (M18AAP) protein (Kumari and Chandra, 2015), (3) predict natural products with antimalarial bioactivity (Egieyeh et al., 2018), (4) screen inhibitors and noninhibitor targeting *P. falciparum* intraerythrocytic cycle (Subramaniam et al., 2011), and also (5) predict that molecules which will block the malarial parasite's ion pump, PfATP4 (Rio et al., 2017).

Therefore, in this context, the authors have applied the principles of ML to form a chemoinformatics model to screen inhibitor of a PfCDPK4 protein essential for sexual reproduction (microgamete formation) and transmission of the *P. falciparum* parasite into mosquitoes (Billker et al., 2004; Solyakov et al., 2011; Tewari et al., 2010). Thus, to build ML-based chemoinformatics and validate the efficiency of the predictive model to screen inhibitor of PfCDPK4 protein the current research article is divided into following three sections (1) Materials and method (2) Results and Discussion and (3) Conclusion. The materials and method section of the research paper describes the dataset as well as provides details about the methods involved in building an ML-based predictive model.

While the Results and Discussion section explains the results obtained in the making and testing of different classifier based-model using different statistical model evaluators. The results obtained show that the predictive model made using J48 classifier is useful in predicting active PfCDPK4 inhibitor from an independent chemical molecule dataset and also outperform other ML-based chemoinformatics models developed for the screening of antimalarial molecules. Moreover, the conclusion section states the importance of our proposed computational predictive model in enhancing the hit-rate of novel antimalarial drug and also discusses the future scope of the present model in antimalarial drug discovery program. Fig. 1 represents a flow diagram describing the making of the computational predictive model for facilitating the rapid screening of antimalarial drugs.

2. Materials and methods

The material and method section describes the following (1) the HTS bioassay data (AID: 1159588), (2) the data processing strategies and (3) the ML classifiers employed in the making of a classifier based model to predict active PfCDPK4 inhibitors from AID: 1159588 bioassay dataset. Moreover, the section also explains different statistical assessors to assess the accuracy of the predictive model to identify active PfCDPK4 protein inhibitor molecules.

2.1. Dataset source

The bioassay dataset AID-1159588 consisted of approximately 13500 cell-active molecules screened against the CDPK4 protein of *P. falciparum* (CDPK4/PF3D7_0717500/XP_001349078.1). The biochemical screening of ~13500 chemical molecules was performed to test *P. falciparum* protein kinases inhibitors. The AID-1159588 dataset was obtained from PubChem bioassay repository (NCBI, 2016). The confirmatory bioassay AID-1159588 tested 55 potent inhibitors of pf-CDPK4 protein (active molecule) and 13396 non-inhibitors of pf-CDPK4 protein (inactive molecules). The Structure Data Format (SDF) of the entire chemical molecules (active and inactive molecules) present in AID: 1159588 bioassay dataset were obtained from PubChem Substance repository (NCBI, 2016).

2.2. Molecular descriptors generation and data pre-processing

The SDF file of the active, as well as inactive chemical molecules obtained from the biochemical screening of *P. falciparum* CDPK4 protein, were fragmented into smaller SDF file using a Perl script (SplitSDFiles) present in MayaChemTool (Sud, 2016). The splitting of the large SDF files of both inactive and active molecules was performed since the memory available in PowerMV can be only utilized to generate molecular descriptors from smaller SDF file of the chemical molecule. The PowerMV is a favorite tool for the generation of the molecular descriptor, molecular similarity search and statistical analysis (Liu et al., 2005). In total 179 two dimensional molecular descriptors (attributes) of each instance (chemical molecule) of both inactive and active chemical molecules were made using PowerMV. Out of 179 molecular descriptors, eight descriptors were categorized based on the chemical property while twenty-four descriptors were classified based on weighted burden numbers and

the rest one forty-seven descriptors belonged to pharmacophore fingerprint. Each molecular descriptor Comma Separated Value (CSV) file of both inactive and active chemical molecule was combined into a lone CSV file. The last column of the individual combined molecular descriptors CSV file was appended with an outcome dependent attribute labeled as "Class." Based on the results of the AID-1159588 bioassay a nominal value "active" or "inactive" for the dependent variable attribute was added for each active and inactive chemical compound.

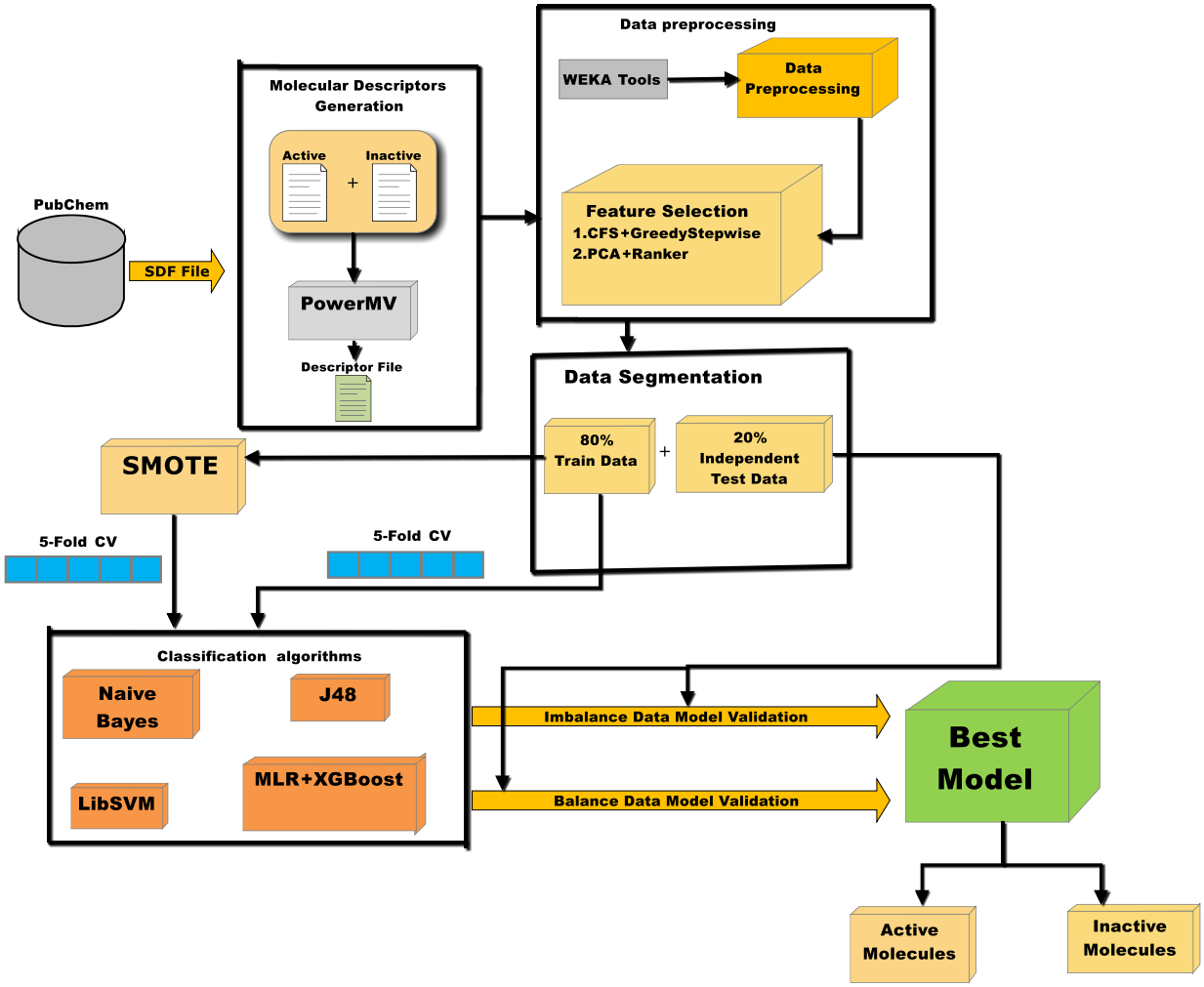

Fig. 1: Making of a classifier based model to accurately screen active inhibitors of PfCDPK4 protein from a given chemical molecule dataset

2.3. Data pre-processing

2.3.1. Data preparation

The single combined CSV molecular descriptor file was preprocessed to remove the noninformative attributes. The attributes having 0's or 1's bit strings throughout the dataset were extracted using the unsupervised "RemoveUseless" attribute filter of Weka software tool. Weka is java based data mining software for data pre-processing, clustering, classification and visualization (Bouckaert et al., 2010). Upon removal of noninformative features, the total number of features for our dataset was reduced to 154 attributes. Furthermore, duplicates instances from the molecular descriptor file were removed using the unsupervised instance filter "RemoveDuplicates" of Weka.

2.3.2. Dimensionality reduction

The feature selection technique plays a pivotal role in constructing an ML-based predictive model with higher interpretability, shorter training time and lower complexity. The Feature selection technique selects the best subset of features from a given set of features to form a predictive model with lower variance and higher accuracy. The main idea behind implementing a feature selection technique is to decrease the dimensionality of the data by taking out features that are redundant or irrelevant and do not contribute significantly to enhance the performance of the model. In the present study, the descriptor file of the inactive and active molecules consisted of 154 features. Therefore, the authors employed the attribute selection methods accessible in the Weka software tool to reduce the dimensionality of the generated chemical descriptor data. In Weka, the feature selection is performed

using both attribute evaluator and search method. Since the dataset understudy have nominal value for the dependent variable (class) and both nominal and binary value for an independent set of variables (attributes), therefore the authors selected two sets of feature selection algorithm. In Weka, each set consists of one feature selection algorithm from attribute evaluator and other from search method. Therefore, in this context, dimensionality reduction of the present dataset is performed by using both CfsSubsetEval algorithms of the attribute evaluator in conjunction with the GreedyStepwise algorithm of the Search method and Principal Components Analysis (PCA) algorithm of the attribute evaluator works in conjunction with the Ranker method of the search method. The CfsSubsetEval select the best subsets of attributes that have a higher value of correlation with the dependent variable (class) and comparatively lower intercorrelation between the independent variable (attributes). While the GreedyStepwise algorithm performs either backward or forward search through the space of attribute subset. The process ends when the deletion or addition of the attributes leads to a decline in the performance of the model. Correspondingly, in PCA the reduction in the dimensionality of the data is made possible by selecting sufficient eigenvectors which account for a small percentage of variance in the original data-default 0.95 (95 %). While the Ranker search method rank attribute based on their performance and is used in combination with PCA algorithm present in the "attribute evaluator" of Weka. Two sets of the dataset were created using the attributes obtained from the two set of feature selection algorithms. Both the dataset with different amount of attributes but the same amount of active and inactive samples was subjected to Synthetic Minority Over-Sampling Technique (SMOTE) for class balancing since the target classes in both the dataset were imbalanced.

2.4. Class balancing using synthetic minority over-sampling technique (SMOTE) algorithm

A dataset is considered imbalanced when the target classes in a classification problem are unequal in number. The AID-115483 bioassay dataset consists of two class's namely inactive and active chemical molecule. The AID-11583 bioassay dataset is highly imbalance as the ratio of active to inactive molecule is 0.0041 therefore SMOTE was used to oversample the minority class (active molecule) by generating artificial instances from the minority class rather than merely making copies of the minority instances from the AID-115483 dataset. The SMOTE algorithm performs the oversampling of the minority class by randomly selecting a random sample from the minority class and selects k nearest neighbor using a distance measure. In this case, five nearest neighbors were selected (k=5) for generating minority synthetic samples. A new synthetic sample is made by multiplying the difference between one of the k neighbor instances and the selected data point (vector) by a random number say "x" which ranges between 0 and 1 (Chawla et al., 2002; Han et al., 2005). In the present study, we have used 0.5, and the result of the same is added to the selected feature vector (data point) to create a new synthetic data point (instance). The final datasets upon implementation of the synthetic minority class oversampling consisted of equal number instances from both the categories (active and inactive).

2.5. Partitioning of dataset and cross-validation

Each set of balanced datasets were divided into 80 % training and 20 % independent validation set. The 80 % training dataset of each set was subjected to 5-fold training-cum-cross validation. The validation of the model was performed using the 20 % independent test data obtained from each dataset.

2.6. Classification algorithms for model building

Classification is a technique where we apply ML algorithms to classify or categorize new data into a given set of classes. In this study, we have used four different state-of-art classification algorithms namely Naïve Bayes (NB), LibSVM, J48, and MLR-XGBoost for building a predictive model to screen active inhibitor compound of PfCDPK4 protein from the 20 % independent test dataset.

The principle of the Bayesian theorem is applied to the Naïve Bayes algorithm, with an assumption that every feature of a given dataset is conditionally independent of each other (Friedman et al., 1997). The LIBSVM is an integrated software tool that implements an SMO-type algorithm for kernelized Support Vector Machine (SVM) supporting regression and classification (Chang and Lin, 2011). The J48 algorithm is a widely used supervised learning algorithm available in Weka for the construction of decision trees from a given labeled dataset. The algorithm initially calculates the entropy (information gain) of the entire set of attributes of the given dataset. Further, the attribute having the least entropy (i.e., highest information gain) is selected as the nonterminal node for splitting the dataset into subsets. Accordingly, the algorithm continues to recur on a subset of attributes to build a decision tree where the nonterminal node represents the splitting point (decision tree node) and a terminal node (leaf node) where all the instances of the subset of independent variable belong to a particular class label (Quinlan, 1993). eXtreme Gradient Boosting (XGBoost) is an implementation of gradient boosting decision tree algorithm. Gradient boosting is an ML technique for solving supervised classification and regression problem. The gradient boosting tree algorithm produces a predictive model in the form of tree ensembles. The tree ensemble is an ensemble of models generated using a set of Classification and Regression Tree (CART). In boosting the ensemble, the model is built by optimizing the training loss and

regularization of different models until no further improvement in the model performance can be made (Chen and Guestrin, 2016).

2.7. Model performance assessment

The performance of various predictive model build on LibSVM, NB, J48 and XGBoost algorithms were assessed using different statistical performance evaluators available in the Weka software tool. The performance evaluations for all predictive models were performed using 20 % independent test dataset unseen by the trained models. The datasets prepared using different feature selection methods were imbalanced. Therefore, the influence of SMOTE on model performance was also evaluated.

Recall or True Positive Rate (TPR) estimates the percentage of True Positives (TP) (Powers, 2011) (i.e., correctly classified positives instances) from the total number of actual positives samples (i.e., False Negative (FN)+True Positive (TP)) and is determined as follows:

$$\frac{TP}{TP+FN} \quad (1)$$

In our case, the recall evaluates the ability of the trained predictive models to correctly classify inhibitors of PfCDPK protein (TP) from 20 % independent test data. Moreover, the proportion of False Positives (FP) (i.e., False Positive Rate (FPR)) obtained from a given population of negative instances [True Negative (TN) + False Positives (FP)] is calculated as follow:

$$\frac{FP}{TN+FP} \quad (2)$$

In this regard, in the usual case, a predictive model with lower FPR can correctly classify TP (inhibitors of PfCDPK) with higher accuracy when compared to a model with higher FPR. Precision refers to the proportion of correctly classified positive instances (TP) from the total number of retrieved positive instances (TP + FP) and is determined as follow (Powers, 2011):

$$\frac{TP}{TP+FP} \quad (3)$$

Therefore, in our case, a model with a higher number of FP instances will have a lower chance to screen true inhibitors of PfCDPK protein from a given independent test data. Moreover, specificity another statistical parameter determines the competency of the model to classify TN instances from a given dataset correctly and is defined as follow:

$$\frac{TN}{TN+FP} \quad (4)$$

In the present case, the model with higher specificity has the higher competency to correctly classify non-inhibitors of PfCDPK protein (TN) from inhibitor chemical molecules (TP). Additionally, the accuracy which determines the ability of the predictive classification model to classify TP and TN instances correctly is an important parameter to help determine the ability of our generated predictive model to accurately discriminate between active and inactive inhibitor chemical molecule of PfCDPK protein from a given independent test data and is determined as follow:

$$(TP + TN)/(TP + TN + FP + FN) \quad (5)$$

The ideal value for accuracy for a predictive model is 1. Therefore, in our study, an ideal predictive model is the one that can accurately classify inhibitors and a noninhibitory molecule of PfCDPK protein as TP and TN, respectively from any given independent chemical test dataset. Moreover, the Area under the Curve (AUC) is a statistical model evaluator who evaluates the consistency of the classification model to predict positive instances (Powers, 2011), i.e., in our case inhibitors of PfCDPK protein from the given dataset. The AUC curve of the model is plotted by plotting the FPR and TPR value of every instance of a dataset in the x-axis and y-axis, respectively. An AUC value of 1is considered as an ideal value for a predictive model. Therefore, in the present case, a model which has a higher AUC value is deemed to be reliable in screening positive sample (i.e., true inhibitors of PfCDPK protein) from any given dataset.

2.8. Selection of best FS method and two sample unpaired t-test

Firstly, the feature selection method which gave better value for different statistical evaluators when tested on models built using four state-of-art classification algorithms (NB, LibSVM, J48, and XGBoost) was selected. Lastly, the statistical significance of SMOTE on dataset prepared using the above chosen FS method was evaluated using two-sample unpaired t-test (Student, 1908; Barbara, 2008). The two sample unpaired t-test was employed to find the statistical significance of sensitivity (recall) in a different predictive model built on balanced and imbalanced data (i.e., with or without SMOTE algorithm).

3. Results and discussion

3.1. Molecular descriptor generation and classification model generation

The confirmatory AID-1159588 bioassay dataset consisted of 55 active, and 13396 inactive molecules screened against CDPK protein of *Plasmodium falciparum*. The SDF of all the inactive and active molecules were downloaded from the PubChem Substance repository, and molecular descriptor dataset was prepared using PowerMV (a molecular descriptor generator software tool) (NCBI, 2016). Firstly 179 2D molecular descriptors were generated

and upon further data pre-processing the number of features (molecular descriptors) was reduced to 154. Since all the features do not contribute significantly in the enhancement of model performance, therefore feature selection technique was employed to define the subset of features that help significantly to build a predictive model with a higher sensitivity in screening active inhibitors of CDPK protein of *P. falciparum*. As per the nature of the dependent attribute and independent variable two set of feature selection method was employed to generate two daughter datasets. The daughter dataset created using CfsSubsetEval, and GreedyStepwise algorithm consisted of eight features including a dependent attribute labeled as "class" to represent the two classes (active and inactive) of the chemical molecule.

Similarly, another daughter dataset consisting of 109 features including the dependent variable attribute (i.e., class) and was created using Principal Components Analysis (PCA) algorithm in conjunction with the Ranker method. Each of the dataset prepared using different FS methods was randomly partitioned into 20 % independent test and 80 % training data. Since the number of instances representing the active class of molecule in the training datasets was far lower as compared to the inactive class of molecule, therefore SMOTE a class balancing algorithm was implemented to equal the number of instances of each class (active and inactive molecule) in the 80 % training dataset.

3.2. Model performance evaluation

Four state-of-art classification algorithms were used to build classification models using Five-fold cross-validation on the training datasets (with and without SMOTE) generated using two FS method. The best model for each classifier was selected based upon their performance using various statistical evaluators namely accuracy, precision, recall, sensitivity and FPR upon testing on 20 % independent test dataset. The model performances of different classifier based models are shown in Table 1.

Since the predictive models were trained and built using both imbalanced and balanced dataset therefore in the present case the statistical evaluator "accuracy" alone cannot be sufficient to determine the effectiveness of the generated predictive models. Thus, the role of other model performance statistical evaluators such as Recall, precision, specificity, and FPR are considered pivotal in determining the effectiveness of the classification model to screen true positive (i.e., inhibitors of PfCDPK protein) instances from any given independent test data.

Table 1: Tabulate the performance of different classifier based models generated using data obtained using different feature selection techniques

Feature Selection Methods	SMOTE	Classifier	ACC	Recall	Precision	Specificity	ROC Area	TP Rate	FP Rate	Confusion Matrix			
										TP	FN	FP	TN
CFS Subset Evaluator + Greedy Stepwise	Not Using SMOTE	Naive Bayes	94.7515	1.000	0.066	0.95	0.993	1.000	0.053	8	0	113	2032
		libSVM	99.6284	0.000	0.000	1.00	0.500	0.000	0.000	0	8	0	2145
		J48	99.6284	0.000	0.000	1.00	0.500	0.000	0.000	0	8	0	2145
		MLR-XGboost	99.582	0.000	0.000	0.99	0.499	0.000	0.000	0	8	1	2144
	Using SMOTE	Naive Bayes	99.6749	0.375	0.600	0.99	0.998	0.375	0.001	3	5	2	2143
		libSVM	94.1013	1.000	0.059	0.98	0.970	1.000	0.059	8	0	127	2118
		J48	97.5848	1.000	0.133	0.97	0.999	1.000	0.024	8	0	52	2093
		MLR-XGboost	96.2843	1.000	0.091	0.96	0.997	1.000	0.037	8	0	80	2065
Principal Component Analysis + Ranker Method	Not Using SMOTE	Naive Bayes	91.7789	0.571	0.022	0.92	0.816	0.571	0.081	4	3	174	1972
		libSVM	99.7213	0.143	1.000	1.00	0.571	0.143	0.000	1	6	0	2146
		J48	99.7213	0.286	0.667	0.99	0.655	0.286	0.000	2	5	1	2145
		MLR+XGboost	99.6749	0.143	0.500	0.99	0.570	0.143	0.000	1	6	1	2145
	Using SMOTE		99.1175	0.857	0.250	0.99	0.993	0.857	0.008	6	1	18	2128
		libSVM	90.8964	1.000	0.034	0.91	0.954	1.000	0.091	7	0	196	1950
		J48	99.8078	0.999	0.997	0.99	0.999	0.857	0.003	6	1	7	2139
		MLR+XGboost	98.3279	0.714	0.128	0.98	0.987	0.714	0.016	5	2	34	2112

The recall is the capability of the classification model to detect TP instances from any given number of positive instances available in an independent test dataset. As tabulated in Table 1 and graphically represented in Fig. 2 and Fig. 3, respectively the recall for different classifiers based models generated using unbalanced dataset (without SMOTE) is nearly zero and just contrary are the results of recall of models made using balanced dataset (with SMOTE) where the recall value is 1 or almost equal to 1. Therefore, the use of SMOTE on datasets generated using both FS method enhanced the capability of the model to screen TP from a given number of actually positive instances.

Similarly, another statistical evaluator namely precision that calculates the proportion of TP instances from a given number of retrieved positive instances and specificity which determine the effectiveness of the model to screen TN accurately from a given number of negative instances were used to assess model performance. In this context, a model with higher precision and specificity will be useful in classifying TP instances from TN instances with higher accuracy. In this context, model based on J48 classifier trained on balanced dataset prepared using PCA in conjunction with Ranker FS method was found to be more effective in discriminating a true inhibitor of PfCDPK protein (TP) from a non-inhibitor of PfCDPK protein (TN) with a precision

value of 0.997 and specificity value of 0.99 as shown in Table 1, and pictorially represented in Fig. 3.

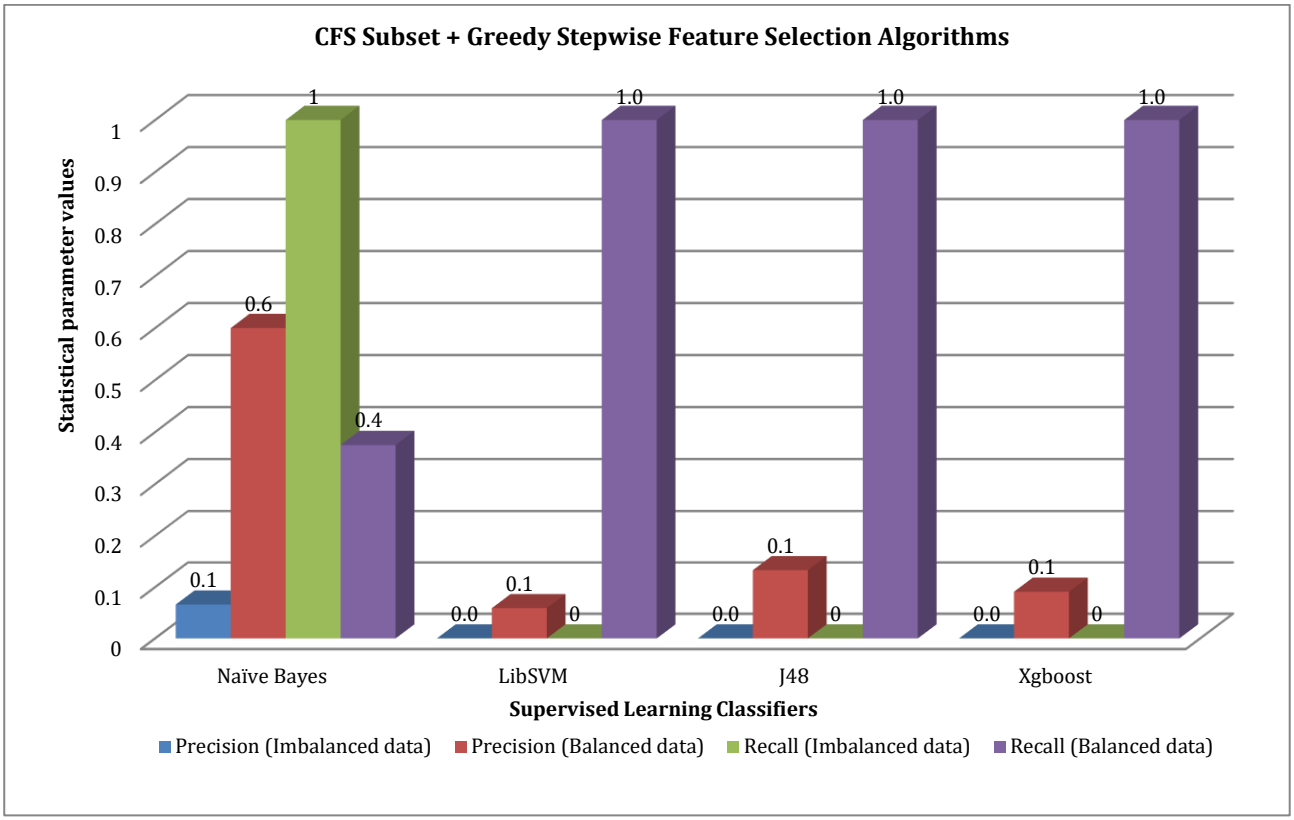

Fig. 2: Comparative statistics evaluation of the four state-of-art classifier based models built using both imbalanced and balanced datasets (The original imbalanced data was generated using CFS Subset Evaluator in association with Greedy Stepwise Feature selection algorithms and that was later balanced using SMOTE algorithm)

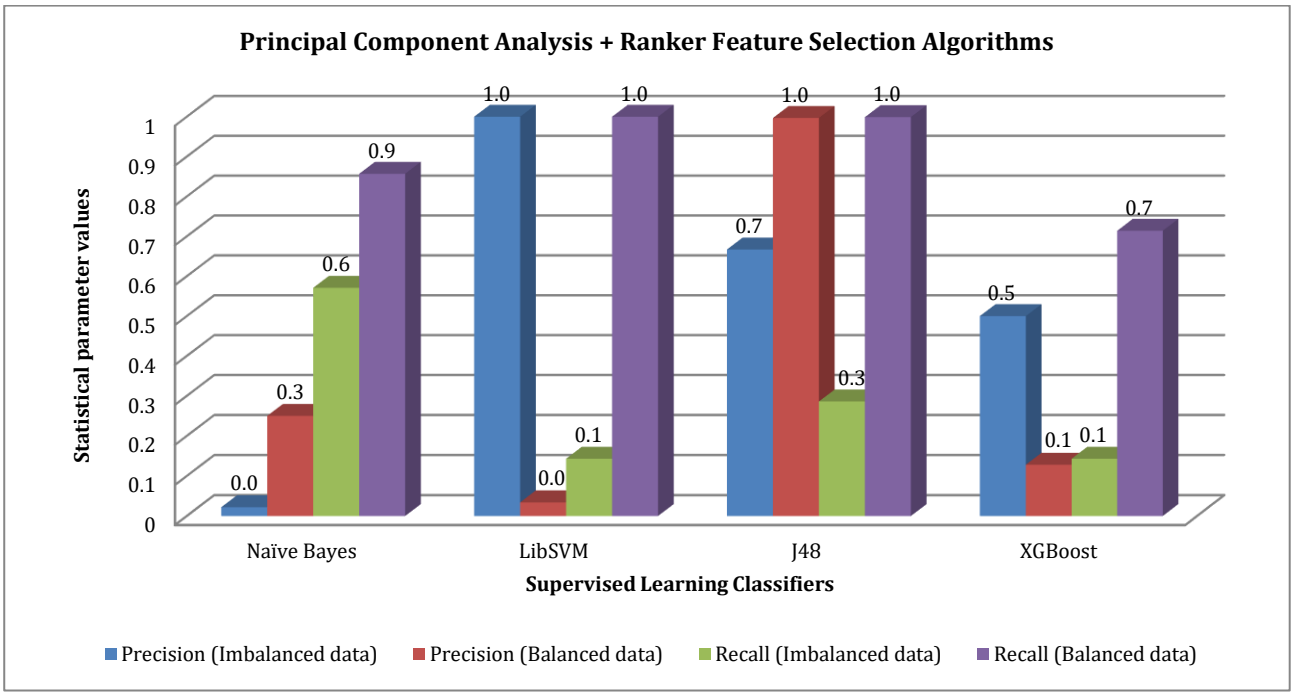

Fig. 3: Comparative statistics evaluation of the four state-of-art classifier based models built on both imbalanced and balanced data (The original imbalanced data was generated using PCA in association with Ranker Feature selection algorithms and that was later balanced using SMOTE algorithm)

Additionally, the FPR that determines the proportion of FP instances from a set of predicted negative instances was found to be 0.003 for J48 classifier based predictive model which is close to an ideal value, i.e., "0" as shown in Table 1. The AUC values obtained from the ROC plot of NB, LibSVM, J48, and XGBoost based classifier model showed higher values i.e., 99 to 100 % for model trained using balanced datasets as compared to the AUC value obtained for above mentioned classifier based models trained using unbalanced dataset as shown in Fig. 4 and Fig. 5, respectively. Therefore, the

classifier based model when trained using balanced dataset has higher reliability in predicting TP instances (inhibitors of PfCDPK protein) from any library of chemical molecules).

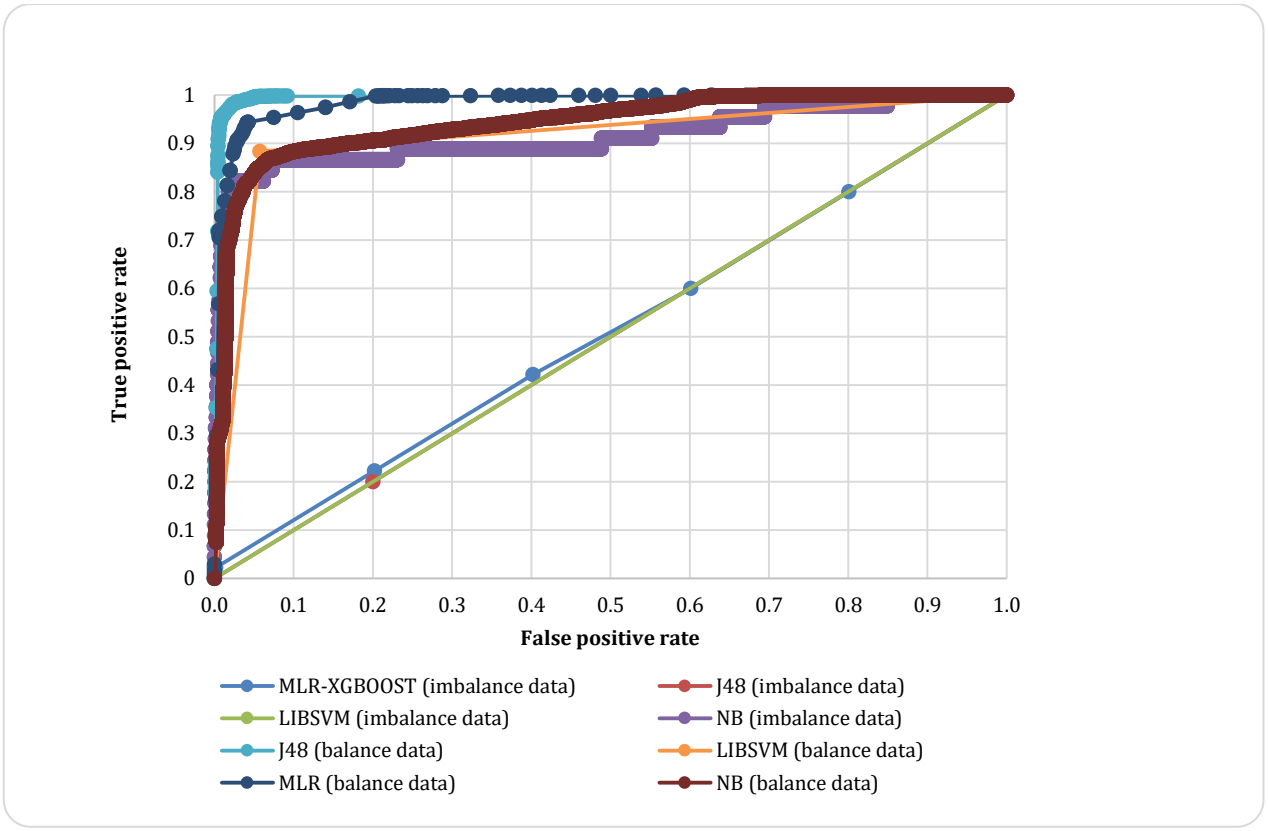

Fig. 4: Comparative ROC plot of the four state-of-art classifier based models built on both imbalanced and balanced data (The original imbalanced data was generated using CFS Subset Evaluator in association with Greedy Stepwise Feature selection algorithms and that was later balanced using SMOTE algorithm)

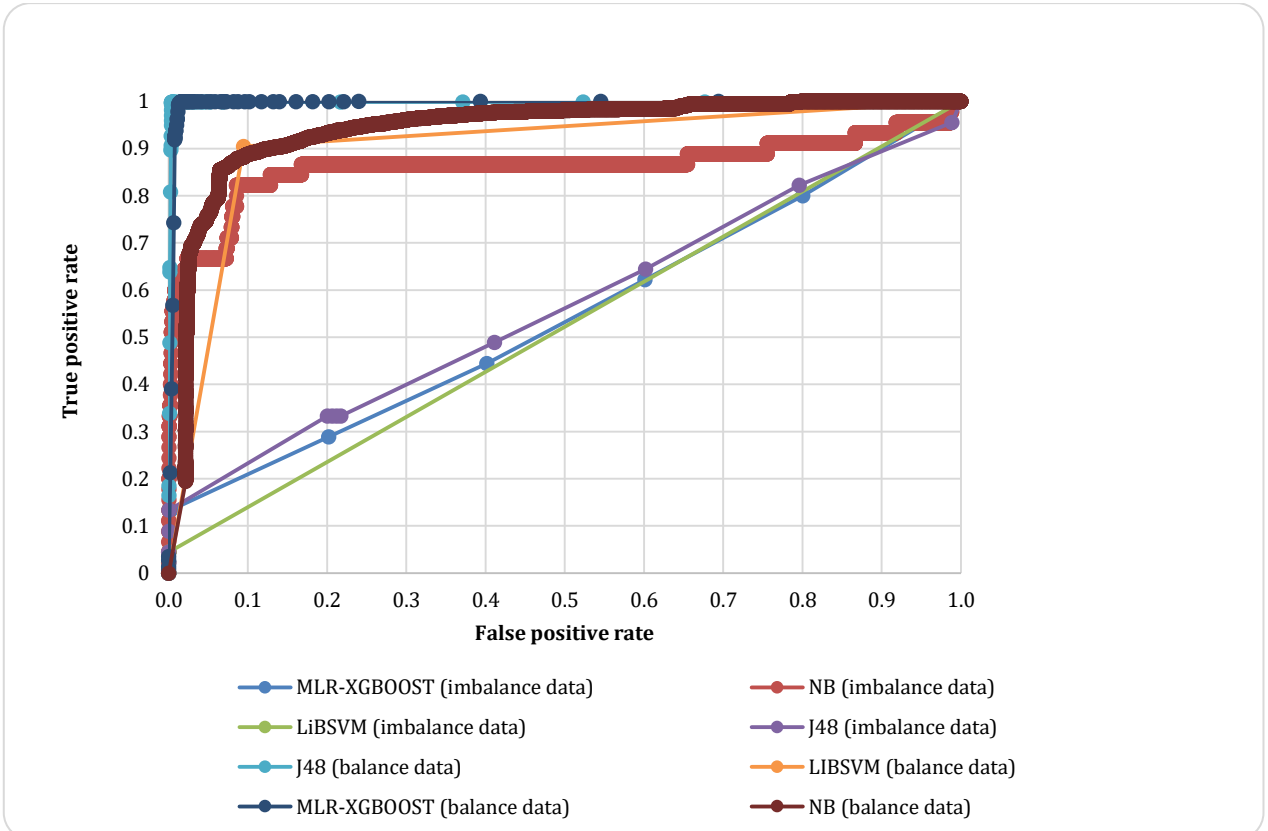

Fig. 5: Comparative ROC plot of the four state-of-art classifiers based models built on both imbalanced and balanced data (The original imbalanced data was generated using PCA in association with Ranker Feature selection algorithms and that was later balanced using SMOTE algorithm)

Accuracy another statistical evaluator for model performance defined as the proportion of correctly predicted TP and TN instances from the total number of predicted positive and negative instances. The J48 classifier based predictive model which was trained on a dataset generated using PCA in conjunction with Ranker method and later class balanced using SMOTE showed better capability in predicting TP and TN samples from the total number of predicted samples. The J48 based predictive model showed a higher value of accuracy, i.e., 0.998, when compared to other classifiers based predictive model generated using dataset, created using PCA and ranker method and balanced using SMOTE. The accuracy value for various classifier based classification model is shown in Table 1. Therefore, based upon the results of the statistical evaluation the J48 classifier based predictive model trained and built on a dataset generated using PCA in conjunction with ranker based FS method was found to be a useful model for screening PfCDPK protein inhibitor molecule from an independent test dataset.

Further, the statistical significance of the application of SMOTE on classifier based models trained on a dataset generated using PCA in conjunction with ranker based FS method was evaluated using Two sample unpaired t-test. The statistical evaluator "recall" was used by the authors to check the effect of SMOTE on the four state-of-art classifier based predictive models. Mean, standard error, standard deviation, and significance value were calculated for the precision values of all the 4 classifier based predictive models trained on both imbalanced (without SMOTE) and balanced (with SMOTE) datasets generated using PCA in conjunction with ranker FS method and further tested on 20 % independent test dataset are shown in Table 2.

The significance value of 0.0025 was obtained when the results of precision were compared for all classifier based models trained and built on the imbalanced and balanced dataset.

Table 2: Unpaired samples T-test for recall was performed between model built using balance (with smote) and imbalanced (without smote) dataset generated using PCA and ranker feature selection (FS) method

Algorithm	Unpaired Differences							
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		T	Df	Sig. (2-tailed)
SMOTE and WITHOUT SMOTE	0.60675	0.244	0.122	0.30861	0.90489	4.9798	6	0.0025

The significance value obtained show that the precision results obtained by all the four state-of-art classifiers based model made using balanced dataset is statistically significant since the calculated significance values are lower than 0.05. Thus it can be proposed that the classifier based models when trained using balanced dataset has a higher efficiency in predicting TP instances (i.e., inhibitors of PfCDPK protein) from any library of chemical molecules.

3.3. A comparative study with other ML-based antimalarial predictive model

The average accuracy and ability of the proposed J48 classifier based predictive model to screen true antimalarial molecule are comparatively higher than any other ML-based chemoinformatics model as shown in Table 3.

Even though different ML-based models were tested on different dataset but still the overall potency to screen true positives from a given balanced dataset is greater as depicted regarding accuracy and AUC value of our proposed J48 classifier based predictive chemoinformatics model. The reasons for better accuracy and AUC value can be inferred from the fact that the application of both SMOTE (class balancer) and selection of appropriate features using PCA and Ranker feature selection method resulted in the development of an efficient antimalarial chemoinformatics predictive model. Based on our results we argue that the current J48 classifier based classification model will be competent in screening true antimalarial molecules from any given independent test chemical dataset.

4. Conclusion and future scope

The current proposed supervised J48 classifier based predicted model built using attributes selected using PCA in conjunction with ranker method showed better performance in screening TP instances, i.e., PfCDPK inhibitor molecules from an independent test dataset. Recall a statistical evaluator was used to assess the outcome of SMOTE a class balancing algorithm on a different classifier based model built using the dataset prepared using PCA in association with ranker FS method. The performance of the four state-of-art classifiers based model to screen TP instances was enhanced when the model was built using balanced dataset (with SMOTE). The results of recall obtained using SMOTE were found be significant when tested using two-sample unpaired t-test at 95 % confidence interval. The comparative study displaying the performance of different classifier based chemoinformatics model show that our J48 classifier based classification model betters other ML-based antimalarial chemoinformatics models regarding accuracy and AUC value. Therefore, the current suggested J48 classifier based predictive model built on balanced class dataset enables a more specific and rapid screening of novel antimalarial drugs targeted against *P. falciparum* CDPK protein.

Table 3: Performance comparison of different machine learning algorithm based chemoinformatics model for screening antimalarial chemical molecules

Author	ML-algorithm	Number of Active Molecule	Number of Inactive Molecule	Class Balancer	Target	ROC area	Accuracy
Subramaniam et al. (2011)	Support Vector Machine (SVM)	443	560	NA	Inhibitors of *Plasmodium falciparum* proliferation	Model 1 0.88	Model 1 87 %
Jamal et al. (2013)	Random Forest (RF)	22396	197741	Cost-Sensitive Classifier	Apicoplast inhibition in the malarial parasite *Plasmodium falciparum*	0.71	76.27 %
Kumari and Chandra (2015)	Random Forest (RF)	3498	287,235	Cost-Sensitive Classifier	Aspartyl aminopeptidase (M18AAP) of *Plasmodium falciparum*	0.86	97.3 %
Bharti and Lynn (2017)	Random Forest (RF)	18126	220632	Down sampling of the data (random sampling has been done to bigger class)	Apicoplast inhibition in the malarial parasite *Plasmodium falciparum*	0.92	88 %
Dixit and Singla (2017)	NA	1173 (AID-504850) 1391(AID-504848)	344 (AID-504850) 240 (AID-504848)	NA	Apicoplast inhibition in the malarial parasite *Plasmodium falciparum*	NA	81.4 %
Egieyeh et al. (2018)	Sequential Minimization Optimization (SMO)	347	808	The Weka "meta-CostSensitiveClassifier" + Synthetic Minority Over-sampling Technique (SMOTE)	Antiplasmodial	0.86	85.9 %
Current research	J48	55	13396	Synthetic Minority Over-sampling Technique (SMOTE) algorithm	*P. falciparum* Calcium-Dependent Protein Kinase4 (PfCDPK4)	0.99	99.8 %

Acknowledgment

The Authors are grateful to the Dean of Faculty of Computing and Information Technology, King Abdulaziz University to provide an excellent platform for conducting various machine learning experiments.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Barbara F (2008). High-yield behavioural science (high-yield series). Lippincott Williams and Wilkins, Hagerstown, USA.

Bharti DR and Lynn AM (2017). QSAR based predictive modeling for anti-malarial molecules. Bioinformation, 13(5): 154-159. https://doi.org/10.6026/97320630013154 **PMid:28690382 PMCid:PMC5498782**

Billker O, Dechamps S, Tewari R, Wenig G, Franke-Fayard B, and Brinkmann V (2004). Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite. Cell, 117(4): 503-514. https://doi.org/10.1016/S0092-8674(04)00449-0

Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P, and Witten IH (2010). WEKA-Experiences with a java open-source project. Journal of Machine Learning Research, 11: 2533-2541.

Burrows JN, van Huijsduijnen RH, Möhrle JJ, Oeuvray C, and Wells TN (2013). Designing the next generation of medicines for malaria control and eradication. Malaria Journal, 12: 187. https://doi.org/10.1186/1475-2875-12-187 **PMid:23742293 PMCid:PMC3685552**

Chang CC and Lin CJ (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3): 27. https://doi.org/10.1145/1961189.1961199

Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

Chen T and Guestrin C (2016). XGBoost: A scalable tree boosting system. In the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, USA: 785-794. https://doi.org/10.1145/2939672.2939785

Dixit S and Singla D (2017). CAPi: Computational model for apicoplast inhibitors prediction against plasmodium parasite. Current Computer-Aided Drug Design, 13(4): 303-310. https://doi.org/10.2174/1573409913666170301121110 **PMid:28260517**

Dua VK, Dev V, Phookan S, Gupta NC, Sharma VP, and Subbarao SK (2003). Multi-drug resistant Plasmodium falciparum malaria in Assam, India: Timing of recurrence and anti-malarial drug concentrations in whole blood. The American Journal of Tropical Medicine and Hygiene, 69(5): 555-557. https://doi.org/10.4269/ajtmh.2003.69.555 **PMid:14695096**

Egieyeh S, Syce J, Malan SF, and Christoffels A (2018). Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. PloS One, 13(9): e0204644. https://doi.org/10.1371/journal.pone.0204644 **PMid:30265702 PMCid:PMC6161899**

Friedman N, Geiger D, and Goldszmidt M (1997). Bayesian network classifiers. Machine Learning, 29(2-3): 131-163. https://doi.org/10.1023/A:1007465528199

Han H, Wang WY, and Mao BH (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In the International Conference on Intelligent Computing, Springer, Nanchang, China: 878-887. https://doi.org/10.1007/11538059_91

Jamal S, Periwal V, and Scaria V (2013). Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics, 14: 55. https://doi.org/10.1186/1471-2105-14-55 **PMid:23419172 PMCid:PMC3599641**

Kumari M and Chandra S (2015). In silico prediction of anti–malarial hit molecules based on machine learning methods. International Journal of Computational Biology and Drug Design, 8(1): 40-53. https://doi.org/10.1504/IJCBDD.2015.068783 **PMid:25869318**

Liu K, Feng J, and Young SS (2005). PowerMV: A software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. Journal of Chemical Information and Modeling, 45(2): 515-522. https://doi.org/10.1021/ci049847v **PMid:15807517**

Mehta SR and Das S (2006). Management of malaria: Recent trends. Journal of Communicable Diseases, 38(2): 130-138.

Melville JL, Burke EK, and Hirst JD (2009). Machine learning in virtual screening. Combinatorial Chemistry and High Throughput Screening, 12(4): 332-343. https://doi.org/10.2174/138620709788167980

NCBI (2016). Biochemical screen of P. falciparum CDPK4. National Center for Biotechnology Information, Bethesda, Maryland, USA.

Powers DM (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1): 37–63.

Quinlan JR (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, USA.

Rio ALD, Llorach-Parés L, Perera-Lluna A, Avila C, Nonell-Canals A, and Sanchez-Martinez M (2017). Machine-learning QSAR model for predicting activity against malaria parasite's ion pump PfATP4 and in silico binding assay validation. Multidisciplinary Digital Publishing Institute Proceedings, 1(6): 652. https://doi.org/10.3390/proceedings1060652

Schierz AC (2009). Virtual screening of bioassay data. Journal of Cheminformatics, 1: 21. https://doi.org/10.1186/1758-2946-1-21 **PMid:20150999 PMCid:PMC2820499**

Solyakov L, Halbert J, Alam MM, Semblat JP, Dorin-Semblat D, Reininger L, and Holland Z (2011). Global kinomic and phospho-proteomic analyses of the human malaria parasite plasmodium falciparum. Nature Communications, 2: 565. https://doi.org/10.1038/ncomms1558 **PMid:22127061**

Student (1908). The probable error of a mean. Biometrika, 6(1): 1–25. https://doi.org/10.1093/biomet/6.1.1

Subramaniam S, Mehrotra M, and Gupta D (2011). Support vector machine based classification model for screening plasmodium falciparum proliferation inhibitors and non-inhibitors. Biomedical Engineering and Computational Biology. https://doi.org/10.4137/BECB.S7503

Sud M (2016). MayaChemTools: An open source package for computational drug discovery. Journal of Chemical Information and Modeling, 56(12): 2292-2297. https://doi.org/10.1021/acs.jcim.6b00505 **PMid:28024397**

Tewari R, Straschil U, Bateman A, Böhme U, Cherevach I, Gong P, and Billker O (2010). The systematic functional analysis of Plasmodium protein kinases identifies essential regulators of mosquito transmission. Cell Host and Microbe, 8(4): 377-387. https://doi.org/10.1016/j.chom.2010.09.006 **PMid:20951971 PMCid:PMC2977076**

Trenholme GM and Carson PE (1978). Therapy and prophylaxis of malaria. Journal of the American Medical Association, 240(21): 2293-2295. https://doi.org/10.1001/jama.240.21.2293 **PMid:359850**

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, and Bryant SH (2009). PubChem: A public information system for analyzing bioactivities of small molecules. Nucleic Acids Research, 37(suppl_2): W623-W633. https://doi.org/10.1093/nar/gkp456 **PMid:19498078 PMCid:PMC2703903**

WHO (2017). Malaria. World Health Organization, Geneva, Switzerland. Available online at: https://bit.ly/1fteTok

Wongsrichanalai C, Pickard AL, Wernsdorfer WH, and Meshnick SR (2002). Epidemiology of drug-resistant malaria. The Lancet Infectious Diseases, 2(4): 209-218. https://doi.org/10.1016/S1473-3099(02)00239-6

Wongsrichanalai C, Webster HK, Wimonwattrawatee T, Sookto P, Chuanak N, Thimasarn K, and Wernsdorfer WH (1992). Emergence of multidrug-resistant plasmodium falciparum in Thailand: In vitro tracking. The American Journal of Tropical Medicine and Hygiene, 47(1): 112-116. https://doi.org/10.4269/ajtmh.1992.47.112 **PMid:1636877**

Yang Z, Li C, Miao M, Zhang Z, Sun X, Meng H, and Cui L (2011). Multidrug-resistant genotypes of Plasmodium falciparum, Myanmar. Emerging Infectious Diseases, 17(3): 498-501. https://doi.org/10.3201/eid1703.100870 **PMid:21392443 PMCid:PMC3166001**