Machine Learning Method to Screen Inhibitors of Virulent Transcription Regulator of *Salmonella* Typhi

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

Atif Hassan
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur,
Kharagpur, West Bengal, India

Tabrej Khan
Department of Information Sciences,
Faculty of Computing and Information Technology Rabigh (FCITR)
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—The PhoP regulon, a two-component regulatory system is a well-studied system of *Salmonella enterica* serotype typhi and has proved to play a crucial role in the pathophysiology of typhoid as well as the intercellular survival of the bacterium within host macrophages. The absence of PhoP regulon in the human system makes regulatory proteins of PhoP regulon for target specific for future drug discovery program against multi-drug resistant strains of *Salmonella enterica* serotype typhi. In recent years, high-throughput screening method has proven to be a reliable source of hit finding against various diseases including typhoid. However, the cost and time involved in HTS are of significant concern. Therefore, there is still a need for an expedient method which is also reliable in screening active hits molecules as well as less time consuming and inexpensive. In this regards, the application of machine learning (ML) based chemoinformatics model to perform HTS of drug-like hit molecules against MDR strain of *Salmonella enterica* serotype typhi is the most applicable. In this study, bagging and gradient boosting based ML algorithm was used to build a predictive classification model to perform virtual HTS of active inhibitors of the PhoP regulon of *Salmonella enterica* serotype typhi. The eXtreme Gradient Boosting (XGBoost) based classification model was comparatively accurate and sensitive in classifying active drug-like inhibitors of PhoP regulon of *Salmonella enterica* serotype typhi.**

Keywords*—Typhoid; PhoP regulon, classification model; machine learning (ML) algorithm; eXtreme Gradient Boosting; random forest; sensitivity; accuracy

I. Introduction

Typhoid is an endemic disease of developing nations caused by *Salmonella enterica* serotype typhi. According to recent WHO data, around 21 million people are infected with *Salmonella* Typhi, and approximately 2, 22, 000 people die annually across the globe [1]. The multidrug-resistant (MDR) strain of *S. typhi* has spread rapidly and has become a major endemic problem in South East Asia and Indian subcontinent [1]. Therefore, the target based screening of novel anti-typhoidal compound with a higher potential to destroy MDR strains of *S. typhi* causing MDR typhoid fever is of prime importance.

The two-component (PhoQ-PhoP) regulon is a crucial virulence regulatory system of *S. typhi* regulating the expression of more than 120 different genes involved pathogenicity of *S. typhi* within the host cells [2]. The PhoP regulon consists of an environmental sensor histidine kinase (PhoQ) that in response host defensins (abundant in macrophages), low level of periplasmic Mg^{2+} ions, acidic pH is activated upon autophosphorylation at the conserved histidine residue present in the cytoplasmic domain of PhoQ protein [3]. Consequently, the PhoP a response regulator of the PhoP regulon is phosphorylated at the aspartate residue present at the conserved N-terminal domain of PhoP protein by accepting a phosphate group from PhoQ protein [4]. The phosphorylated PhoP regulates the transcription of corresponding genes involved in the intracellular survival [5]-[6] and virulence of the *S. typhi* within host cells [7]-[9]. The PhoP/PhoQ operon based virulence regulatory system is not present is present only in a bacterial system; therefore the PhoP regulon has gained significance as a potential target for antibacterial drug discovery program.

ML algorithms are robust and fast in dealing with high dimensional data. Since the chemical dataset used for screening of drug-like lead molecules during the earlier stages of drug discovery involves high dimensional data, i.e., comprising a large number of two dimensional and three-dimensional chemical attributes. Therefore, ML-based methods are most appropriate in categorizing inactive and active compounds from a given library of chemical compounds. Chemoinformatics models based on ML algorithms has been suitably applied in the past to screen as well as rank active hit molecules during the lead molecule identification stages of drug discovery and development program. In this regard, Garcia-Sosa et al. 2012 [10] used multivariate logistic regression methods to classify active and inactive drug-like molecules. On the other hand, Korkmaz et al. 2014 [11] used a combination of various feature selection method with Support Vector Machine (SVM) to discriminate active and inactive drug-like molecule from on similar dataset. Further, Korkmaz et al. 2015 [12], proposed a web tool (MLViS) using the best ML-based classification algorithms to screen active drug-like molecule during the early stages of drug discovery protocols. A comparative study to evaluate the performance of SVM and Neural network (NN) based classification model to discriminate between a drug-like and nondrug-like was conducted by Zernov et al. [13] and Byvatov et al. [14]. They both showed that SVM based model performed better in classifying drug-like molecule from the non-drug-like molecule. Similarly, SVM algorithm based classification model was also used to classify inhibitors of cytochrome P450 [15], lymphocyte-specific protein tyrosine kinase [16] and butyrylcholinesterase [17]. On the other hand, other ML algorithms such as k-Nearest Neighbor (KNN) [18], NN [19-20] and Naïve Bayes (NB) [2], were used to classify active inhibitors from non-inhibitor molecules. Likewise, SVM algorithm based predictive model was used by Rathke *et al*. [22], Wassermann *et al*. [23], Jorissen and Gilson [24], and Agarwal et al. 2010 [25] to evaluate chemical compound based on their activity. Similarly, Abdo et al. 2010 [26] and Plewczynski et al. 2009 [27] have applied Random Forest (RF), and Bayesian neural network (BNN) based predictive model for predicting the activity of chemical molecules. Additionally, Harlen et al. 2012 [28] used Random Forest (RF) algorithm to build a predictive model to classify active and inactive chemical molecule against the PhoP regulon of *S. typhi* from the HTS bioassay dataset. The accuracy, sensitivity, and specificity obtained using the RF classifier based model was 81.5%, 87.7%, and 81.5%, respectively. In this context, an improved classification model built using the supervised ML-based algorithms (XGBoost and RF) have been proposed to classify active inhibitors of PhoP transcriptional regulatory system protein (PhoP) with higher accuracy, sensitivity and specificity than proposed and built by Harlen et al. 2012 [28]. Since the number of the active molecule with a potential to inhibit PhoP regulon was less as compared to their inactive counterpart in the AID-1850 dataset, therefore, the dataset was balanced using Synthetic Minority Over-sampling Technique (SMOTE) algorithm prior applying supervised ML algorithm for model building. The basic idea of the proposed model is to build a less expensive and robust predictive classification model which will be potent in screening active inhibitors of PhoP regulon and thus will save time and money for identifying lead molecule during the early stages of anti-tuberculosis drug discovery program. The present original research article is sectioned into four sections: In order to overcome the problem associated with the high-cost experimental screening protocols, the current research work in Section II. Firstly, defines the dataset used for building the chemoinformatic classification model; secondly apply SMOTE algorithm to balance the dataset since the AID-1850 dataset is highly imbalanced, and finally discuss various Classification algorithms namely RF and XGBoost used to construct the current supervised classification model. Section III explains the results of the statistical model performance evaluators for the classification model build using balanced bioassay dataset, and Section IV provide the concluding statements about the proposed classification model as well as the future scope of the proposed model. A pictorial representation of the workflow diagram involved in constructing the supervised classification model for classifying active inhibitors of PhoP protein is summarized and is shown in Fig. 1.

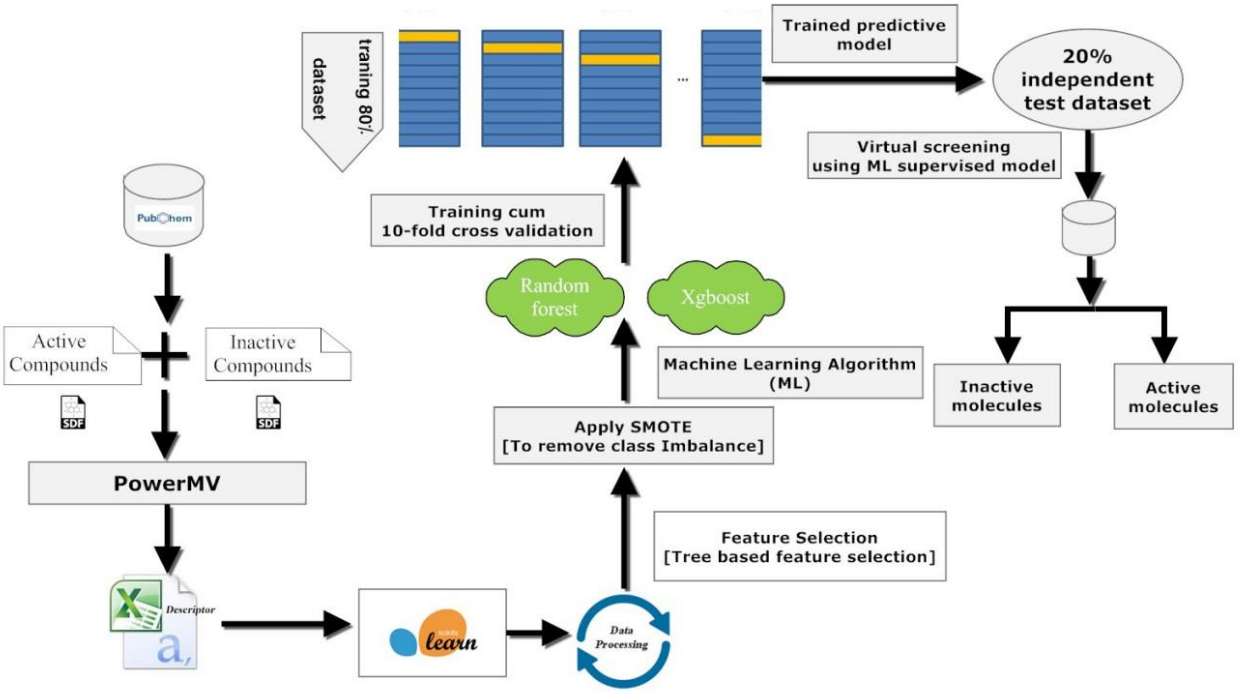
traning 80% dataset
...
Trained predictive model
20% independent test dataset
Virtual screening using ML supervised model
PubChem
Training cum 10-fold cross validation
Random forest
Xgboost
Active Compounds
Inactive Compounds
SDF
SDF
Machine Learning Algorithm (ML)
Inactive molecules
Active molecules
Apply SMOTE [To remove class Imbalance]
PowerMV
Feature Selection [Tree based feature selection]
Descriptor
scikit learn
Data Processing

Fig. 1. A pictorial representation of workflow to illustrate the methods required to build a supervised classification model to screen active inhibitors of PhoP operon proteins.

II. MATERIALS AND METHODS

This part describes the dataset and defines the process to pre-process and balance the dataset. This part also explains the ML algorithm used to build the supervised classification model as well as describes the statistical evaluators used to access the performance of the ML-based classification models.

A. Data Source

The dataset AID=1850 was obtained from the PubChem BioAssay Database of the National Center for Biotechnology Information (https://pubchem.ncbi.nlm.nih.gov/bioassay/1850). A total of 306568 compounds were screened for a compound that inhibits the PhoP operon in *S. typhi.* The compounds based on a percentage of inhibition were classified into the active and inactive molecule. The molecules which showed > 30% inhibition in the confirmatory PhoP dose response assay were considered as active while the molecules which showed less than 30% of inhibition in the dose-response test were deemed to be inactive. Therefore, the trial generated 1021 active inhibitors of PhoP regulon and 305404 inactive compounds in the confirmatory bioassay.

B. Attribute Generation

The Structural-data files (SDF) of both active and inactive compounds from AID-1850 dataset were downloaded from PubChem bioassay dataset [29]. The molecular descriptor file for both active and inactive compounds of the AID-1850 dataset was generated using PowerMV, a Graphical User interface (GUI) based software for molecular descriptor generation and visualization [30]. A Perl script based Mayachemtool[1] was used to split the sizeable structural-data file of the inactive compounds into smaller structural-data files. Using PowerMV, 179 molecular descriptors (attributes) were generated from the chemical-data record for each active and inactive compounds of the AID-1850 dataset. Bit string and continuous calculation method were used create the molecular descriptor file for each active and inactive compounds present in the AID-1850 dataset. A total of 147 molecular descriptors generated based on pharmacophore fingerprinting were represented as a bit string, i.e., 0 and 1. Where the bit string "1" signifies the occurrence of a specific feature/ fragment and "0" represent the absence of that specific fragment/feature. Twenty four weight burden number and eight chemical properties based continuous molecular descriptors were generated using PowerMV. A list of eight property based descriptors namely the number of rotatable bonds, Polar surface area (PSA), XLogP, molecular weight, a molecule containing the toxic group (bad group indicator) and blood-brain indicator represented by 1 and 0. Here the discrete value "1" displays the ability of the molecule to cross the Blood-Brain Barrier (BBB) and while "0" indicates the inability of the compound to pass the BBB. The molecular descriptor file of each active and inactive compounds consisting of 179 descriptors was combined and saved as Comma-separated Value (CSV) file for further processing. An extra column depicting the outcome (bioactivity) of each instance (compound) was appended. The inhibitors of PhoP compounds were given a nominal value "active," and the non-inhibitors of PhoP were labeled as "inactive."

C. Processing of Clinical Dataset

1) Data Preparation

The molecule id column was removed from each matrix, as it does not contribute to the feature list. The combined file CSV consisting of active and inactive molecules was preprocessed to remove the duplicate instances. As a result, 352 duplicate samples including 11 active compound samples were removed from the dataset. A quick count shows that a total of 300992 samples were present with the majority class being the inactive compounds and occupying 98.34% of the sample space whereas the 1010 active compounds being the minority class hold only 1.66% of the total sample space. Further, the final dataset after removal of duplicate instances was subjected to filtration of non-informative attributes to improve the efficiency of the model generated using ML tools [31], [32]. The removal of non-informative attribute reduces the feature space of AID-1850 dataset to 154 attributes. The final list of 154 attributes are enlisted and shown in Table I (Supplementary File).

2) Dimensionality Reduction

Using all features of a given dataset is not an efficient model building process as higher dimensions add to the complexity of the final classifier which leads to longer computation time while unimportant features reduce the model performance/accuracy. Since the feature space of the dataset was 154, therefore a tree-based feature selection module[2] to reduce the dimensionality of dataset to only 43 features listed in Table II (Supplementary File). When an attribute is used as a decision node in a tree, its relative rank/depth can determine how important it concerns the prediction of the target variable. Since the features used at the top of a decision tree affect the final prediction of a large number of input data. Thus, the fraction of samples that they influence can be used to estimate the importance of each feature against one another. By creating some randomized trees and averaging the importance value of each feature, a more robust feature selection model with lower variance can be constructed. A pictorial representation of the scoring based selection of attributes using Tree-based feature selection method is shown in Fig. 2.

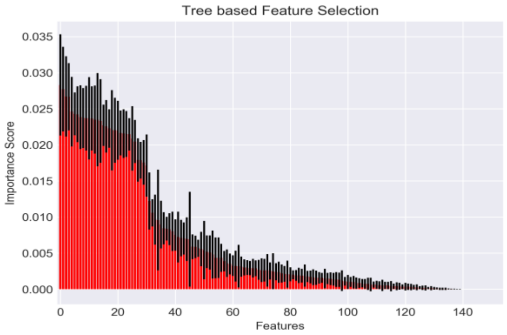

Fig. 2. Illustrates the score obtained by the 154 features of the AID-1850 dataset using Tree-based feature selection module of Scikit-learn package.

[1] http://www.mayachemtools.org/

[2] http://scikitlearn.org/stable/modules/feature_selection.html#tree-based-feature-selection

3) Class Balancing

A dataset is called imbalanced if the numbers of target classes are not nearly equally represented. As the present dataset was highly imbalanced with the majority class taking up 98.34% of the total sample space, SMOTE algorithm was used to balance the AID-1850 dataset by creating synthetic instances from the minority class from the AID-1850 dataset. SMOTE is an oversampling method which creates synthetic instances of the minority class rather than oversampling the class through replacement. Oversampling is done by randomly choosing a minority sample from the given data and finding its k nearest neighbors [33], [34]. In the present study, k equal to 5 was used. A sample is generated on the line segment joining any or all k neighbors by multiplying the difference between the selected feature vector (instance) and its nearest neighbor with a random value in the range of 0-1. In the present study, a random value 0.5 was selected and multiplied by the feature vector under consideration leading to the generation of a new sample. Similar action is performed for all/any neighbors which effectively forces the minority class decision region to become more general. The final dataset, after completing class balancing consisted of 50% active and 50% inactive instances. The pseudo code for generating a synthetic sample is as follows:

Let a feature vector $\overrightarrow{\boldsymbol{a}}$ represents the instance under consideration. Find its k nearest neighbors and select one of them. Let this instance be represented as a feature vector $\overrightarrow{\boldsymbol{b}}$. Then the new sample $\vec{\boldsymbol{c}}$ will be equal to

$$\overrightarrow{\boldsymbol{c}} = \vec{\boldsymbol{a}} + \left(\vec{\boldsymbol{b}} - \vec{\boldsymbol{a}}\right) * \boldsymbol{rand}(\mathbf{0}, \mathbf{1}) \qquad (1)$$

Where, “rand (0, 1)” represents a random value between 0 and 1.

D. Data Partitioning and Cross-validation Procedures

The balanced dataset was segmented into train and test sets as 80% and 20% respectively. The train set would be used to train the model while the test set, data never before seen by the model, would be used for testing its accuracy/performance. The train set was used for k-fold cross-validation; in the present case, k is 10. Thus, in 10-fold cross validation, one fold is used for testing purpose while the rest 9 (k-1) folds are used for training. This process is repeated until all folds have been tested. The average accuracy taken over all folds gives a more reliable measure than a single training and testing phase. This action can also be used to verify if the final, trained model overfits the test set or not.

E. Model Building Algorithm

Classification is the process of segregating a sample, based on its attributes into the given target classes. In this study, two algorithms namely Random forest and XGBoost were used to perform classification where both of which were based on the concepts of decision trees. A decision tree is a flowchart-like tree structure whose internal nodes are attributes which are used to split the tree further based on a threshold and whose leaf nodes are the target classes. The feature selected at each level for further splitting is determined by calculating information gain for each attribute and the one with maximum information gain is selected. Calculation of the information gain for each feature is done by calculating the entropy of all possible values of an attribute and then finding its information gain. Entropy is the amount of homogeneity of a sample while information gain is the difference of entropies before and after a split. The attribute which yields the maximum information gain is chosen as the root node.

$$\boldsymbol{E}(\boldsymbol{X}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{c}} -\boldsymbol{p}_{\boldsymbol{i}} \boldsymbol{log}_{\mathbf{2}} \boldsymbol{p}_{\boldsymbol{i}} \qquad (2)$$

Entropy using the frequency table of one attribute

$$\boldsymbol{E}(\boldsymbol{T}, \boldsymbol{X}) = \sum_{c \in X} \boldsymbol{P}(\boldsymbol{c}) \boldsymbol{E}(\boldsymbol{c}) \qquad (3)$$

Entropy using the frequency table of two attributes

$$\boldsymbol{Gain}(\boldsymbol{T}, \boldsymbol{X}) = \boldsymbol{Entropy}(\boldsymbol{T}) - \boldsymbol{Entropy}(\boldsymbol{T}, \boldsymbol{X}) \qquad (4)$$

Information Gain

1) Random Forest (RF)

A random forest is a forest/collection of decision trees and works on the concept of bagging. Some un-correlated classifiers/learners can be used together to form a better classifier with less variance and reduce overfitting [35]. In a random forest classifier, the training samples are divided into some random subsets based on which decision trees are created. The target class for a particular instance is then selected based on the maximum voting scheme wherein each tree outputs a class, and the one with the highest votes is chosen as the target class for the given sample. This reduces variance in the present model while also decreases overfitting, a problem which usually occurs in decision trees.

2) XGBoost

XGBoost is an ensemble algorithm which is majorly used in kaggle competitions as it provides excellent performance out of the box and has some parameters for tuning. XGBoost is based on the concept of gradient boosting [36]. Gradient Boosting is a technique which uses an ensemble of weak classifiers, models which perform slightly better than random guesses, to create a strong classifier. Here CART (Classification and Regression Trees) was used for preparing an ensemble algorithm for classification. In boosting, a model is built to optimize a differentiable loss function, at runtime. In the next stage/iteration, a new model is developed to optimize the loss function from the previous step further. This process continues until a threshold is reached. In this way, the errors committed by the earlier models can be corrected by the models in the next stage. XGBoost works on the idea of gradient boosting but differs from the fact that it uses regularized models to control overfitting which gives better performance. At the same time, XGBoost or eXtreme Gradient Boosting is called so because it utilizes other computational techniques such as cache access patterns, data compression, etc. to push computational boundaries and achieve state of the art performance regarding speed and accuracy.

$$\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \left[\boldsymbol{g}_i \boldsymbol{f}_{\boldsymbol{t}}(\boldsymbol{x}_i) + \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{h}_i \boldsymbol{f}_{\boldsymbol{t}}{}^{\mathbf{2}}(\boldsymbol{x}_i)\right] + \boldsymbol{\Omega}(\boldsymbol{f}_{\boldsymbol{t}}) \qquad (5)$$

Where, g_i and h_i are inputs defined by

$$\boldsymbol{g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})} \tag{6}$$

$$\boldsymbol{h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})} \tag{7}$$

$f_t(x)$ is given by

$$\boldsymbol{f_t(x) = w_{q(x)}, w \in R^T, q: R^d \to \{1, 2, \dots\dots, T\}} \tag{8}$$

Where w is the vector of scores on the nodes, T is the number of leaves and q is a function assigning each data point to the corresponding leaf, While $\Omega(f_t)$ is the regularizer term, given by

$$\boldsymbol{\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j{}^2} \tag{9}$$

The ML classification algorithm, the data preprocessing and post-processing data analysis are performed in Scikit-learn tool for data mining and analysis. Scikit-learn tool is a state-of-the-art implementation of many supervised and unsupervised ML algorithms written in and for Python. It has an easy-to-use interface and is a go-to choice for exploratory data analysis. It is increasingly used by academicians as it allows for more time on designing of algorithms than their implementation [37]. Due to the exceedingly imbalance characteristics of the AID-1850 data, the Imbalanced-learn tool was used tackle the problem of class imbalance in the AID-1850 dataset. The imbalanced-learn toolkit is an implementation of some popular re-sampling techniques which is useful for datasets with highly imbalanced classes [38]. It is a Python package and is compatible with Scikit-learn tool and can be downloaded from http://contrib.scikit-learn.org/imbalanced-learn/stable/.

F. *Evaluation of Model Performance*

The ML-based predictive model trained using XGBoost, and RF classifier was evaluated using statistical model performance evaluators present in Scikit-learn data mining tool. For two-class problem the 2x2 confusion matrix consists of the following sections: (1) True Positive (TP) in this study, is the active inhibitors of PhoP appropriately classified by classification model as active class; (2) False Positive (FP) actually non-inhibitors of PhoP operon but incorrectly classified as active by the predictive model; (3) True Negative (TN) actual non-inhibitor of PhoP (inactive molecule) correctly classified by the model as non-inhibitor (inactive) and lastly, (4) False Negative (FN) actually inhibitor molecule (active molecule) but incorrectly classified by the classification model as non-inhibitor (inactive). In this context, the True Positive Rate (TPR) is defined as the proportion of TP (i.e., active inhibitors of PhoP regulon predicted correctly by the classification model) from the total population of inhibitors of PhoP regulon and is estimated as TP/TP + FN. Similarly, False Positive Rate (FPR) is defined as the fraction of FP (i.e., erroneously categorized as active inhibitor molecule) when compared to the total number of predicted inactive chemical molecule and the FPR is estimated as FP/FP+TN.

Sensitivity, another model statistical evaluator, represents the capacity of the ML-based predictive model to correctly classify the inhibitors (True Positives) of PhoP regulon from the instances given in AID-1850 dataset and is estimated as (TP/TP+FN)*100. On the other hand, specificity refers to the capability of the ML-based predictive model to classify inactive (non-inhibitor) molecules from AID-1850 dataset correctly and is estimated as (TN/TN+FP)* 100.

Accuracy is another statistical evaluator to measure the ability of the model to correctly classify the TN and TP instances from the total number of predicted instances (TP+TN+FP+FN) and is calculated as

([TP+TN])/ [TP+TN+FP+FN])*100 (10)

The ideal accuracy value for any classification model is one. The Receiver Operating Characteristics (ROC) graph defines the consistency of the model to efficiently discriminate between two classes by Area under the Curve (AUC) value. The AUC value is generated by making a graph between True Positive Rate (TPR) on the Y-axis and False Positive Rate (FPR) on the x-axis. The estimated AUC value is the likelihood that the model will assign a higher score to an arbitrarily selected inhibitor compound than to an arbitrarily picked non-inhibitor compound. The score of AUC value ranges from 0 to 1, therefore the model with a score close or equal to 1 will be considered reliable in predicting active compound from the AID-1850 chemical dataset and vice versa for a model with an AUC value close to zero.

G. *Determination of Statistically Significant Difference between Models*

Estimation of a significant statistical difference between the models generated using XGBoost and RF in predicting active molecule from AID 1850 dataset was determined using two-sample unpaired *t*-test [39]. The accuracy value of each test-fold obtained during the ten-fold training cum-cross validation of XGBoost and RF model were grouped and tested for significant difference using two-sample t-tests at a confidence interval of 95%.

III. Results and Discussion

The HTS dataset AID-1850 was obtained from the publically available bioassay database of PubChem-NCBI. The HTS dataset consisted of 1021 active molecules (inhibitors of PhoP regulon) and 305404 inactive molecules (non-inhibitors of PhoP regulon). The SDF files of the inactive and active molecules were taken from the bioassay database of PubChem. Due to the larger size, the SDF file of the compounds was fragmented into smaller files using MayaChemTool. A CSV file consisting of 179 molecular descriptors for each active and inactive molecule were generated using PowerMV. Finally, each descriptor files of both active and inactive molecule were randomly merged into one CSV file. The Final merged molecular descriptor CSV file of both active and inactive molecule was preprocessed in Scikit-learn platform to remove duplicate instances and noninformative attributes. Since only specific attributes among the total attributes contribute significantly in accurately predicting the desired class. Therefore, in this context Tree-based feature selection module of Scikit-learn package was used to remove noninformative features or attributes from the final merged molecular descriptor file of active and inactive molecules of AID-1850 HTS chemical dataset. The last molecular descriptor file of both active and inactive molecule consisted of 43 attributes (molecular descriptors) contributing most towards model building were screened using tree-based feature selection method. While an outcome column was labeled as “class” having two discrete values “0” and “1”. Here the value “1” denotes an inactive molecule and “0” signifies an active molecule.

The modified molecular descriptor file of both active and inactive molecule with 43 attributes was split into 80% training data and 20% independent test data. Due to the imbalanced nature of the dataset, SMOTE algorithm was applied to generate synthetic instances from the minority class (active molecule) to create a balance between the two classes (i.e., active and inactive molecule) of the dataset. Therefore, the final dataset post SMOTE application consisted of 50% active and 50% inactive instances. The balanced dataset was fragmented into 20% independent test dataset and 80% training dataset. The classification models built using XGBoost and RF algorithms were trained using the 80% training dataset. Furthermore, the classification models trained using 80% training data were tested for their ability to classify active and inactive molecule from the 20% independent test dataset. A comparative performance evaluation of the predictive models with and without SMOTE is tabulated in Table III. All the results tabulated in Table III for each classifier is determined using 20 % independent test data. Due to the imbalanced nature of the dataset, the models tend to be predisposed to the majority class (inactive molecule).The biasness of the models for the majority class can be observed from the results of both FNR and sensitivity. Here, sensitivity reflects the capability of the model to appropriately categorize the True Positive (active molecule) instances from AID-1850 HTS dataset while FNR reflects the prejudiced nature of models for the majority class inactive molecule). In this regard, the models tested on an imbalanced dataset shows a higher rate of FN’s (79.6% for RF and 85.7% for XGBoost) and a lower percentage of sensitivity (20.4% for RF and 14.1% for XGBoost). The above results show that the predictive classification models built and tested on imbalance dataset are biased towards the majority class (inactive molecule). However, contrasting results are obtained post-SMOTE application. The sensitivity and the FNR for the classification models constructed using RF and XGBoost algorithm and tested on balanced dataset show a lower value for FNR (0.7 for RF and 2.4 for XGBoost) and a higher percentage for sensitivity (99.2% for RF and 97.8% for XGBoost) as shown in Table III. The higher percentage of sensitivity for RF-based model show its higher ability to accurately predict active molecule (TP) instances from total positive instance (TP+FN) present in AID-1850 HTS dataset when compared to XGBoost based predictive model. The ability of the predictive system to accurately predict TP post SMOTE application can also be determined by the TPR. The TPR value before the application of SMOTE was 25.4% for RF and 14.1% for XGBoost based predictive models. Subsequently, the TPR value post SMOTE application is 99.2% for RF and 97.6% for XGBoost, based classification models. Further, the accuracy and specificity of the models tested on the imbalanced test data have similar value for each model (i.e., 98.5% accuracy for each model and 99.8% specificity for each model). Furthermore, post SMOTE application the RF-based classification model had similar accuracy and specificity value (99.2%), which was comparatively better than accuracy and specificity, obtained using the XGBoost classifier based predictive model. Since, the RF classifier based classification model achieves higher percentages of specificity and sensitivity in detecting TN (inactive molecules) and TP (active molecules) samples, respectively from a balanced AID-1850 chemical dataset. Therefore, the classification model built using the RF algorithm is considered as in ideal model to screen active PhoP regulon inhibitors from a given balanced chemical HTS dataset. Evaluation of the ability of the classifiers to selectively classify TP from the FN instances present in the balanced dataset is another significant statistical model performance evaluator. In this regard, the AUC value is the probability that classifier based model will give a higher score to an arbitrarily selected positive sample (active molecule) when compared to an arbitrarily selected negative sample (inactive sample). The AUC value for a classification model is calculated by plotting a ROC curve between the TPR in the y-axis and FPR in the x-axis. The AUC value for a classification model ranges from 0 to 1. Thus a model with an AUC value close or equal to 1 is considered as an ideal model to selectively choose a positive instance from mixed instances of positive (active molecule) and negative (inactive molecule) instances in a given dataset.

TABLE III. The Performance Evaluation of RF and Xgboost Classifier Based Supervised Classification Models

SMOTE	**Classifier**	**Area Under the Curve (AUC)**	**Accuracy**	**True Positive Rate (%)**	**False Positive Rate (%)**	**True Negative Rate (%)**	**False Negative Rate (%)**	**Specificity (%)**	**Sensitivity (%)**
Not Using SMOTE	Random Forest	0.86	98.533	25.398	0.151	99.849	79.602	99.849	20.398
	XGBoost	0.93	98.525	14.136	0.126	99.874	85.864	99.874	14.136
Using SMOTE	**Random Forest**	**0.99**	**99.237**	**99.243**	**0.768**	**99.231**	**0.757**	**99.232**	**99.243**
	XGBoost	0.97	97.75	97.636	2.138	97.862	2.364	97.862	97.636

Fig. 3 shows the comparative ROC plot of RF and XGBoost classification model built and tested on balanced 20 % independent tests dataset. It can be apparently comprehended from Fig. 3 that the AUC value of RF algorithm-based model which is 0.99 is comparatively higher than that of XGBoost classifier based predictive model. Thus, the RF classifier based predictive model is considered as an efficient model for selectively classifying positive instances (active PhoP inhibitor molecules) from a given balanced chemical dataset. The statistically significant difference between the RF and XGBoost based classification models in accurately predicting the active and inactive molecule was determined using two sample Unpaired t-test, and the result is tabulated in Table IV. An exceptionally statistically significant two-tailed P value is less than 0.0001 was obtained when the mean of ten-fold accuracy values of the RF-based trained model was compared with XGBoost classification trained model at 95 % confidence interval.

The present proposed classification model based on RF classifier is more sensitive and accurate in classifying active PhoP inhibitors molecules from the AID-1850 dataset as compared to classification model proposed by Kaur et al. 2016 as observed from the present findings tabulated in Table V. The accuracy AUC, specificity, and sensitivity of the current RF based model is higher as compared to accuracy AUC, specificity, and sensitivity obtained by the base classifiers used by Kaur et al. 2016 to build a predictive model to classify inhibitors of PhoP operon from the AIS-1850 dataset.

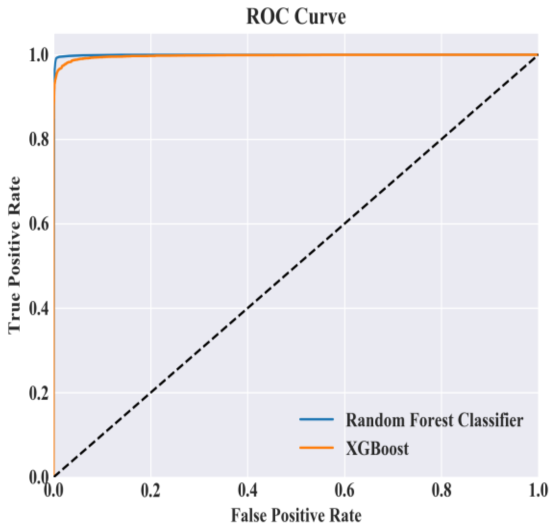

Fig. 3. Comparative ROC plot of RF and XGBoost algorithm based supervised classification tool on balanced dataset.

Moreover, the FNR of the present study is lower as compared to the FNR obtained by predictive models proposed by Kaur et al. 2016. Therefore, the current model based on RF-based classifier built and tested on the balanced dataset is far more superior in screening a real positive (active PhoP inhibitor molecule) from a given AID-1850 dataset.

TABLE IV. Two-Sample Unpaired T-Test To Determine Significant Difference Between RF and Xgboost Classifier Based Predictive Model

Algorithms	**Paired Differences**							
				95% Confidence Interval of the Difference				
	Mean	**Std. Deviation**	**Std. Error Mean**	***Lower***	***Upper***	**t**	**Df**	**Sig. (2-tailed P-value)**
RF and XGBoost	0.0166900490	.00196	0.00062	0.0156290243	0.0177510737	33.0478	18	<0.0001

TABLE V. Comparative Performance Evaluation of the RF Classifier Based Supervised Classification Model with Another Predictive Model for Screening Active Inhibitors of Phop Regulon Protein

ML classifier based classification model	**AUC**	**Accuracy**	**FNR**	**sensitivity**	**specificity**
RF (Kaur et al. 2016)	91.5	81.5	12.3	87.7	81.5
RF (Hassan et al. 2018)	**0.99**	**99.2**	**0.8**	**99.2**	**99.2**

Here, AUC=Area Under the Curve; FNR=False Negative Rate

IV. Conclusion and Future Scope

In the current study, ML algorithm is used to build a supervised classification model to classify active PhoP inhibitor molecules from the balanced AID-1850 HTS dataset. The capability of the predictive model to distinguish between the active and inactive classes of the AID-1850 dataset was determined by specific attributes selected using the tree-based feature selection module of Scikit learn package. The final 43 descriptors based dataset was processed using SMOTE algorithm to remove the class imbalance present in the AID-1850 dataset. Several statistical assessors were used to assess the performance of RF and XGBoost classifier based classification model in screening true inhibitors of PhoP regulon protein from a given dataset. The comparative performance evaluation of both XGBoost and RF classifier based predictive model revealed that RF classifier based model showed better ability to predict active PhoP inhibitors from the preprocessed balanced AID-1850 dataset. Moreover, the present RF classifier based model is far more superior in predicting active inhibitors of PhoP regulon protein from AID-1850 when compared to the model proposed by Kaur et al. 2016. Therefore, the present predictive model will be a step forward in screening novel drug-like inhibitors of PhoP a virulent two-component regulatory system of *S. typhi*. Moreover, in future, a web-based real-time predictive system will be built based on the results of the present model to efficiently classify active inhibitors of PhoP operon protein from the sizeable molecular library of molecules from various chemical databases.

ACKNOWLEDGMENT

We are thankful to the Dean Prof. Omar Barukab of the Faculty of Computing and Information Technology Rabigh (FCITR) of King Abdulaziz University, Jeddah for providing the computational facility to perform current experiments.

REFERENCES

[1] S. Now and V. P. Editions, "Release of the 2017 Global Vaccine Action Plan Reports Past Meetings / Workshops," no. November, 2017.

[2] Kato and E. A. Groisman, "The PhoQ/PhoP regulatory network of Salmonella enterica.," Adv. Exp. Med. Biol., vol. 631, pp. 7–21, 2008.

[3] D. Shin and E. A. Groisman, "Signal-dependent Binding of the Response Regulators PhoP and PmrA to Their Target Promoters in Vivo," vol. 280, no. 6, pp. 4089–4094, 2005.

[4] M. E. Castelli, A. Cauerhff, M. Amongero, F. C. Soncini, and E. G. Vescovi, "The H box-harboring domain is key to the function of the Salmonella enterica PhoQ Mg2+-sensor in the recognition of its partner PhoP.," J. Biol. Chem., vol. 278, no. 26, pp. 23579–85, Jun. 2003.

[5] L. R. Prost and S. I. Miller, "The Salmonellae PhoQ sensor: mechanisms of detection of phagosome signals.," Cell. Microbiol., vol. 10, no. 3, pp. 576–582, Mar. 2008.

[6] B. Blanc-Potard and E. A. Groisman, "The Salmonella selC locus contains a pathogenicity island mediating intramacrophage survival.," EMBO J., vol. 16, no. 17, pp. 5376–5385, Sep. 1997.

[7] W. S. Pulkkinen and S. I. Miller, "A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein.," J. Bacteriol., vol. 173, no. 1, pp. 86–93, Jan. 1991.

[8] E. Garcia Vescovi, F. C. Soncini, and E. A. Groisman, "Mg2+ as an extracellular signal: environmental regulation of Salmonella virulence.," Cell, vol. 84, no. 1, pp. 165–174, Jan. 1996.

[9] L. Guo, K. B. Lim, C. M. Poduje, M. Daniel, J. S. Gunn, M. Hackett, and S. I. Miller, "Lipid A acylation and bacterial resistance against vertebrate antimicrobial peptides.," Cell, vol. 95, no. 2, pp. 189–198, Oct. 1998.

[10] T. Garcia-Sosa, M. Oja, C. Hetenyi, and U. Maran, "DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties.," J. Chem. Inf. Model., vol. 52, no. 8, pp. 2165–2180, Aug. 2012.

[11] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "Drug/nondrug classification using Support Vector Machines with various feature selection strategies.," Comput. Methods Programs Biomed., vol. 117, no. 2, pp. 51–60, Nov. 2014.

[12] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development.," PLoS One, vol. 10, no. 4, p. e0124600, 2015.

[13] V. V Zernov, K. V Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V Pletnev, "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions.," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 2048–2056, 2003.

[14] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 1882–1889, 2003.

[15] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P. W. Lee, and Y. Tang, "Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers.," J. Chem. Inf. Model., vol. 51, no. 5, pp. 996–1011, May 2011.

[16] C. Y. Liew, X. H. Ma, X. Liu, and C. W. Yap, "SVM model for virtual screening of Lck inhibitors.," J. Chem. Inf. Model., vol. 49, no. 4, pp. 877–885, Apr. 2009.

[17] J. Fang, R. Yang, L. Gao, D. Zhou, S. Yang, A.-L. Liu, and G. Du, "Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery.," J. Chem. Inf. Model., vol. 53, no. 11, pp. 3009–3020, Nov. 2013.

[18] D. W. Miller, "Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning," J. Chem. Inf. Comput. Sci., vol. 41, no. 1, pp. 168–175, Jan. 2001.

[19] Ajay, W. P. Walters, and M. A. Murcko, "Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules?," J. Med. Chem., vol. 41, no. 18, pp. 3314–3324, Aug. 1998.

[20] J. Sadowski and H. Kubinyi, "A scoring scheme for discriminating between drugs and nondrugs.," J. Med. Chem., vol. 41, no. 18, pp. 3325–3329, Aug. 1998.

[21] H. Sun, "A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing.," J. Med. Chem., vol. 48, no. 12, pp. 4031–4039, Jun. 2005.

[22] F. Rathke, K. Hansen, U. Brefeld, and K.-R. Müller, "StructRank: A New Approach for Ligand-Based Virtual Screening," J. Chem. Inf. Model., vol. 51, no. 1, pp. 83–92, Jan. 2011.

[23] M. Wassermann, H. Geppert, and J. Bajorath, "Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors.," J. Chem. Inf. Model., vol. 49, no. 3, pp. 582–592, Mar. 2009.

[24] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine.," J. Chem. Inf. Model., vol. 45, no. 3, pp. 549–561, 2005.

[25] S. Agarwal, D. Dugar, and S. Sengupta, "Ranking chemical structures for drug discovery: a new machine learning approach.," J. Chem. Inf. Model., vol. 50, no. 5, pp. 716–731, May 2010.

[26] Abdo, B. Chen, C. Mueller, N. Salim, and P. Willett, "Ligand-based virtual screening using Bayesian networks.," J. Chem. Inf. Model., vol. 50, no. 6, pp. 1012–1020, Jun. 2010.

[27] Plewczynski D, Grotthuss MV, Rychlewski L, Ginalski K. Virtual high throughput screening using combined random forest and flexible docking. Comb Chem High T Scr. 2009;12: 484–489.

[28] H. Kaur, M. Ahmad, and V. Scaria, "Computational Analysis and In silico Predictive Modeling for Inhibitors of PhoP Regulon in S. typhi on High-Throughput Screening Bioassay Dataset," Interdiscip. Sci. Comput. Life Sci., vol. 8, no. 1, pp. 95–101, 2016.

[29] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," Nucleic Acids Res., vol. 37, no. Web Server issue, pp. W623–W633, Jul. 2009.

[30] K. Liu, J. Feng, and S. S. Young, "PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation," J. Chem. Inf. Model., vol. 45, no. 2, pp. 515–522, Mar. 2005.

[31] H. Kaur, R. Chauhan, and S. K. Wasan, "A Bayesian Network Model for Probability Estimation," in Encyclopedia of Information Science and Technology, Third Edition, IGI Global, pp. 1551–1558.

[32] H. Kaur, R. Chauhan, M. A. Alam, S. Aljunid, and M. Salleh, "SpaGRID: A Spatial Grid Framework for High Dimensional Medical Databases BT - Hybrid Artificial Intelligent Systems," 2012, pp. 690–704.

[33] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[34] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning BT - Advances in Intelligent Computing," 2005, pp. 878–887.

[35] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[36] T. Chen and C. Guestrin, "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, pp. 785–794.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J.

Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2012.

[38] G. Lemaitre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning,” vol. 18, pp. 1–5, 2016.

[39] Fadem, Barbara (2008). High-Yield Behavioral Science (High-Yield Series). Hagerstwon, MD: Lippincott Williams & Wilkins. ISBN 0-7817-8258-9; "The Probable Error of a Mean" (PDF). Biometrika. 6 (1): 1–25. 1908.

SUPPLEMENTARY FILE

TABLE I. ALL FEATURES SORTED ACCORDING TO IMPORTANCE

WBN_GC_L_1.00
WBN_GC_L_0.25
WBN_GC_L_0.50
WBN_LP_H_0.25
WBN_LP_L_0.75
WBN_GC_L_0.75
WBN_GC_H_0.75
WBN_GC_H_0.25
WBN_LP_L_0.25
WBN_EN_H_0.25
WBN_EN_H_0.50
WBN_LP_H_0.75
WBN_LP_H_0.50
WBN_EN_L_0.75
WBN_LP_L_1.00
WBN_EN_L_0.25
BadGroup
WBN_GC_H_0.50
WBN_EN_L_0.50
NumRot
WBN_LP_L_0.50
WBN_LP_H_1.00
WBN_EN_H_1.00
WBN_EN_L_1.00
WBN_EN_H_0.75
PSA
WBN_GC_H_1.00
XLogP
MW
NumHBA
NumHBD
BBB
ARC_05_ARC
HBA_03_ARC
POS_04_ARC
ARC_03_ARC
ARC_06_ARC
ARC_04_ARC
HBA_05_ARC
HBA_04_ARC
HBD_02_ARC
ARC_01_ARC
ARC_02_ARC
POS_07_ARC
POS_05_ARC
POS_06_HYP
HBA_06_ARC
HBD_03_ARC
HBA_07_ARC
ARC_07_ARC
POS_02_ARC
POS_06_ARC
HBD_05_ARC
POS_03_ARC
HBD_04_ARC
ARC_03_HYP
NEG_02_NEG
HBD_06_ARC
POS_04_POS
HBA_06_HBA
HBD_07_ARC

NEG_04_NEG
HBA_04_HYP
ARC_05_HYP
POS_03_POS
HBA_03_HBA
ARC_06_HYP
HBD_05_HBD
HBD_05_HBA
POS_04_HBA
ARC_07_HYP
NEG_04_ARC
NEG_05_HYP
ARC_02_HYP
HBA_06_HYP
POS_02_HYP
POS_05_POS
POS_05_HBA
HBD_07_HBA
NEG_02_ARC
ARC_04_HYP
POS_07_HBA
HBD_04_HBA
POS_04_HBD
HBA_07_HBA
POS_07_HBD
HBA_02_HYP
HBA_05_HBA
HBA_07_HYP
HBD_04_HBD
HBD_06_HBA
POS_06_HBA
POS_05_HYP
HBA_03_HYP
HBD_06_HBD
HBA_04_HBA
HBD_07_HBD
POS_05_HBD
HBD_03_HYP
NEG_05_ARC
HYP_01_HYP
NEG_05_HBA
POS_03_HYP
NEG_07_ARC
NEG_06_ARC
HBA_05_HYP
HBD_07_HYP
HYP_03_HYP
NEG_04_HBD
HBD_06_HYP
NEG_03_HBD
POS_07_HYP
NEG_07_HYP
NEG_03_ARC
NEG_07_HBD
POS_03_HBA
POS_04_HYP
HBD_02_HYP
HBD_05_HYP
NEG_06_HBA
POS_06_HBD
HYP_05_HYP

POS_06_POS
HYP_02_HYP
NEG_07_HBA
HBD_04_HYP
HYP_04_HYP
POS_07_POS
NEG_02_HYP
HBD_03_HBD
HYP_07_HYP
NEG_04_HBA
NEG_01_NEG
NEG_03_HBA
NEG_06_HBD
NEG_04_POS
NEG_05_NEG
NEG_03_HYP

NEG_06_HYP
NEG_06_POS
HYP_06_HYP
NEG_01_HBD
NEG_05_POS
HBD_03_HBA
NEG_02_HBD
POS_03_HBD
NEG_03_POS
NEG_07_NEG
NEG_06_NEG
NEG_05_HBD
POS_02_HBD
NEG_03_NEG
NEG_04_HYP
NEG_07_POS

TABLE II. TOP 43 FEATURES SORTED ACCORDING TO IMPORTANCE

WBN_GC_L_1.00
WBN_GC_L_0.25
WBN_GC_L_0.50
WBN_LP_H_0.25
WBN_LP_L_0.75
WBN_GC_L_0.75
WBN_GC_H_0.75
WBN_GC_H_0.25
WBN_LP_L_0.25
WBN_EN_H_0.25
WBN_EN_H_0.50
WBN_LP_H_0.75
WBN_LP_H_0.50
WBN_EN_L_0.75
WBN_LP_L_1.00
WBN_EN_L_0.25
BadGroup
WBN_GC_H_0.50
WBN_EN_L_0.50
NumRot
WBN_LP_L_0.50
WBN_LP_H_1.00

WBN_EN_H_1.00
WBN_EN_L_1.00
WBN_EN_H_0.75
PSA
WBN_GC_H_1.00
XLogP
MW
NumHBA
NumHBD
BBB
ARC_05_ARC
HBA_03_ARC
POS_04_ARC
ARC_03_ARC
ARC_06_ARC
ARC_04_ARC
HBA_05_ARC
HBA_04_ARC
HBD_02_ARC
ARC_01_ARC
ARC_02_ARC