An Improved Machine Learning Approach to Enhance the Predictive Accuracy for Screening Potential Active USP1/UAF1 Inhibitors

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

Ahmed Hamza Osman
Department of Information System
Faculty of Computing and Information Technology at Rabigh,
King Abdulaziz University
Jeddah, Saudi Arab

***Abstract*—DNA repair mechanism is an important mechanism employed by the cancerous cell to survive the DNA damages induced during uncontrolled proliferation of cell and anti-cancer drug treatments. In this context, the Ubiquitin-Specific Proteases (USP1) in complex with Ubiquitin Associated Factor 1(UAF1) plays a key role in the survival of cancerous cell by DNA repair mechanism. Thus, this put forth USP1/UAF1 complex as a striking anti-cancer target for screening of anti-cancer molecule. The current research is aimed to improve the classification accuracy of the existing bioactivity predictive chemoinformatics model for screening potential active USP1/UAF1 inhibitors from high-throughput screening data. The current study employed feature selection method to extract key molecular descriptors from the publicly available high-throughput screening dataset of small molecules that were used to screen active USP1/UAF1 complex inhibitors. This study proposes an improved predictive machine learning approach using the feature selection technique and two class Linear Discriminant Technique (LDA) algorithm to accurately predict the active novel USP1/UAF1 inhibitor compounds.**

Keywords*—Ubiquitinases; DNA repair mechanism; anti USP1/UAF1 molecule; High-throughput Dataset; Feature Selection and Discriminant Technique; Chemoinformatic Model; Classification accuracy; T-test

I. INTRODUCTION

Deubiquitinases (DUBs) are a specific group of enzymatic proteins that aid the process of deubiquitination on targeted proteins [1-2]. Recent findings have highlighted the role of deubiquitinases as oncogenes, due to their involvement in DNA damage repair mechanism leading to the survival of actively replicating cancerous cells [3 to 5]. The DUBs are broadly categorized into five families and the Ubiquitin-specific proteases (USPs) family constitutes of the largest number of different USPs. Among the many members of USPs, the USP1 is the most studied deubiquitinases due to its involvement in various type of carcinomas. Cancerous cell undergoes DNA damage during targeted anti-cancer drug therapy and uncontrolled rapid cell proliferation [6-7]. This leads to dependencies of the cancerous cell upon DNA damage repair mechanism for their continuous proliferation and persistence [8]. The upregulated USP1 in cancerous cell promotes the DNA damage repair pathway enabling the survival and proliferation of the DNA damaged cancerous cell [3-4]. Therefore, inhibition of DNA repair pathway is currently a very eminent anti-cancer strategy [9-10]. Past Studies from various researchers have shown that DNA repair mechanism of USP1 is carried out in the association of a cofactor UAF1 (USP1 associated factor 1), that controls the enzyme activity of deubiquitinases [11-12]. The association of a cofactor UAF1 induces a conformational change in the active site of USP1 thereby increasing the deubiquitinases activity naturally by stabilizing it [13]. It is be noted that upon treatment of DNA targeted drug make the cancerous cell dependent on DNA repair mechanism of USP1 for survival, therefore a combined therapy of UAF1 inhibitor with DNA-damaging therapeutic molecule will enhance the therapeutic efficacy of the therapy against cancer. Thus, this makes the USP1/UAF1 complex a potential anti-cancer target for the exploration of molecules having anti deubiquitinases activity [14]. In this context, the University of Delaware and the NIH Chemical Genomics Center developed a miniaturized quantitative high-throughput screen assay to identify small molecule having anti USP1 activity from the NIH Molecular Libraries Small Molecule Repository (MLSMR) from PubChem [15]. Considering the significance of identifying more inhibitors to USP1/UAF1 complex a chemoinformatic classification model was built using the predictive capacities of machine learning approaches [16]. The machine learning based predictive computational model proposed by Wahi et al. 2015 has a potential to screen potentially active inhibitors of USP1. However, the accuracy of base classifier (random forest) selected for building the predictive model had a sensitivity of 79.44 %, specificity of 81.36 % and an accuracy of 81.35 %, which is presumably low for an efficient and rigorous chemoinformatic predictive model. The objective of the present study was to develop a more rigorous chemoinformatic model for predicting potentially active USP1 inhibitors with high accuracy, sensitivity, and specificity. The proposed method is a hybrid technique based on feature selection technique and discriminant algorithm for active USP1 inhibitor molecule prediction. The proposed classification method seek to increase the accuracy of classifying active USP1 inhibitors from high throughput screens so that genuine hits are optimized using a low-cost large-scale computational virtual screening tool.

The later part of the research article is organized as Sections II present the description of the AID 743255 dataset and an elaborate description of the methodology. In Section III the results of the hybrid technique are discussed. Section IV report the conclusions of the present research work.

II. Material and Methods

A. Bioassay dataset

In the present study, the high throughput screening data set conforming to bioassay identifier AID 743255 was targeted to screen inhibitors of the USP1/UAF1 complex [14]. The dataset comprised of 389,560 compounds and based on their PubChem activity score the compounds were characterized into the active and inactive molecule. The chemical compounds with an activity score of zero were considered inactive (n=369,898) and compounds with a score ranging from 40 to 100 were considered active (n=904). Moreover, the remaining compounds with a score ranging from 0 to 39 were considered unspecific and irrelevant and were not considered for further analysis.

B. Predictive model building

In order to build a Machine learning based predictive tool, a workflow has been built to predict the active USP1 inhibitors from AID 743255 dataset by employing Data Mining Techniques (DMT) for the analysis of high-throughput screen data, and then the result of DMT are extracted to be used as a Knowledge Base for our model to carry out the prediction process. Fig. 1 shows the proposed workflow consisting of (1) Pre-processing of dataset and generation of molecular descriptors; (2) Determination of Best fit descriptors and data segmentation (3) Implementation of classification algorithm, (4) evaluation phase to evaluate the performance and accuracy of the built model using a data mining evaluation technique.

1) Pre-processing of dataset and generation of molecular descriptors

The structural Data format (SDF) files of both the active and inactive compound from bioassay AID 743255 dataset were downloaded from PubChem. Since it was not possible to process the whole SDF file of both active and inactive molecule as a single file, therefore, the SDF files of both the group of molecules were divided into files of smaller sizes by applying the SplitSDFiles present in Mayachem tools [17]. Furthermore, PowerMV a publicly accessible software for descriptor creation and viewing [18] was applied to create two-dimensional molecular descriptors for both the inactive and active compounds of AID 743255 dataset. A total of 179 descriptors were created from the input structural files of compounds using PowerMV of which 8 descriptors were assigned for property descriptor, 24 descriptors were classified under weighted burden numbers and 147 descriptors accounts for pharmacophore fingerprint. The property class of molecular descriptors includes a properties namely Blood-brain barrier (BBB), H-bond acceptors and donors, molecular weight, bad group indicator, the number of rotatable bonds, partition coefficient, and polar surface area.

A group of continuous molecular descriptor based on the burden connectivity matrix namely weighted burden numbers were generated by PowerMV. The burden connectivity matrix considers three important properties namely partial charge, atomic lipophilicity, and electronegativity. Lastly, Pharmacophore fingerprints are descriptors which are expressed as 0 and 1 (binary form) and the grouping of atoms and group are based on biosteric principles such that the atoms and groups having similar activity are grouped together in a specific group (class). Pharmacophore fingerprint descriptors in PowerMV are classified into six major groups that include, ring systems containing aromatic and hydrophobic centers, hydrogen bond donors and acceptors, and positively and negatively charged atoms or groups.

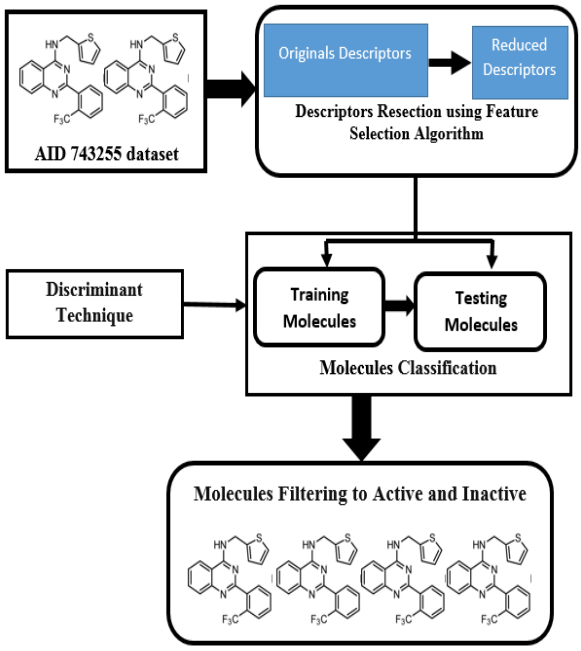

Fig. 1. Proposed workflow for the generation of predictive machine learning based chemoinformatic model

2) Determination of Best fit descriptors and data segmentation

Feature selection (FS) is a technique to pre-process the dataset so that repeated descriptors can be removed and include descriptors which are of relevance in model building. Employing feature selection strategy will not only reduce the dimensionality of the dataset but also will enhance the computational process of the model by reducing the computation time to analyze large data and eliminate the noise from the dataset [19]. The feature selection algorithm explores all set of combinations of molecular descriptors from the dataset and brings forth features which contribute most towards the construction of an efficient classification model [20]. Feature selection algorithm employs search method in combination with a feature evalautor method [21]. This experiment conducted to differentiate the active and inactive molecular from the AID 743255 dataset. Feature selection method was applied first as a feature reduction to reduce the number of the molecules descriptors. Only the number of the extracted descriptors using feature selection algorithm is considered as significant features. Then, the AID 743255 dataset was divided into 10 parts as 10-folds cross validation. Each part had certain molecules (active and inactive). The

experiments were run 10 times with nine parts of these groups as training dataset and one part as a testing dataset.

3) Implementation of classification algorithms

In chemoinformatics, machine learning approaches have been used in the past to build predictive chemoinformatics model from sets of known compounds and predict biological activities of the unknown molecule [22-25]. In this study, to categorize and classify the active USP1 inhibitor molecules from the inactive molecules from the AID 743255 dataset, the two class Linear Discriminant Algorithm (LDA) was applied on the training and testing data. Two class LDA have previously been successfully applied to classify cancer based on gene expression data and has been reviewed as one of the important tools for chemoinformatics classification studies [26-27]. The basic concept of two class LDA is to calculate a linear transformation that helps in binary classification of the data set and the classification is executed in the transformed area formed based on some distance metrics namely euclidean distance as proposed by Fisher, 1936 [28] and shown using the following equations:

Assume that we have a set of "n" number of molecules with f dimensional features (attribute) x_1, x_2, ⋯, x_n (where $x_i = (x_{i1}, \cdots, x_{if})$) classified into two classes, C_1 and C_2. Here C_1= Active molecule and C_2= Inactive molecule. Scatter matrices for given two classes (active and inactive molecule) is shown below:

$$S_{i=} \sum_{X \in C_i} (X - \bar{X}_i)(X - \bar{X}_i\acute{}) \quad (1)$$

Here $\bar{X}_i = \frac{1}{n_i} \sum_{X \in C_i} X$ and n_i is the total number of molecules present in C_i. Therefore, the total scatter matrix for Intra-class is represented as:

$$\sum\nolimits_w = S_1 + S_2 = \sum_i \sum_{X \in C_i} (X - \bar{X}_i)(X - \bar{X}_i\acute{}) \quad (2)$$

The scatter matrix for inter-class is calculated as

$$\sum_b = \sum_{i=1}^{n} (X - \bar{X}_i)(X - \bar{X}_i\acute{}) \quad (3)$$

Where $\bar{X}_i$ is the mean for each class and $\bar{X}$ is total mean vector given by $\bar{X} = \sum_{i=1}^{n} \bar{X}_i$ [29]. Rayleigh coefficient, for the proposed sample, is defined as the ratio of the determinant for the inter and intraclass scatter matrix. For the maximum utilization of Rayleigh coefficient fisher recommended the use linear transformation (Φ):

$$\mathcal{J}(\phi) = \frac{|\phi^t \hat{\sum}_b \phi|}{|\phi^t \hat{\sum}_w \phi|} \quad (4)$$

Equation (3) can be answered as an eigenvalue problem provided $\hat{\sum}_w$ is non-singular, and subsequently ϕ is calculated using the matrix $\hat{\sum}_w^{-1} \hat{\sum}_b$. of eigenvectors.

After transformation Φ is calculated, the classification of the dataset into specific classes is performed within the transformed space based on Euclidean distance and cosine measure, respectively. The equations 5 and 6 represents the calculation of distance using Euclidean distance and cosine measure, respectively:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (5)$$

$$d(x, y) = 1 - \frac{\sum_i (x_i - y_i)^2}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}. \quad (6)$$

Once instance z is initiated, the instance z is classified to

$$= arg_k min\, d(z\phi, \bar{x}_k) \quad (7)$$

Here $\bar{x}_k$ is the centroid of the k[th] class.

The pseudo code for the execution of LDA algorithm for processing AID 743255 dataset is illustrated in Figure 2.

```
Algorithm Linear Discriminant Analysis (LDA)
  Input:
  //A Set X of Active and Inactive molecules{with descriptors}
  //The LDA is "trained"
      LDA test = new LDA(data, group, true);
  //Now we will try to classify new data
    double[] testData = { testi };

System.out.println("Predicted group: " + test.predict(testData));
  //Let's have a look at the values of the discriminat functions

double[] values = test.getDiscriminantFunctionValues(testData);
    For(int i = 0; i < values.length; i++){
       System.out.println("Discriminant function " + (i+1)
            + ": " + values[i]);
    End for
```

Fig. 2. Pseudo code for the execution of LDA algorithm in AID 743255 dataset

In all the cross-validation experiment applied on the dataset, accuracy result, and area under the curve (AUC) were computed. The classifying accuracy calculated using the standard classification equation:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FP}) + (\text{TP} + \text{FN})} \quad (8)$$

Where,

True Positive (TP): The active molecules correctly categorized as active; False Positive (FP): The inactive molecules that were incorrectly classified as active.; True Negative (TN): The inactive molecules correctly classified as inactive; False Negative (FN): The active molecules incorrectly classified as inactive molecules.

SPSS Clementine tool was used to perform the experimentation and the analysis of results. SPSS Clementine tool is an SPSS enterprise-strength data mining workbench. The Clementine tool is used by business organizations to enhance the client and people relations by performing a thorough consideration and analysis of data [30].

III. Experimental Results and Discussion

A. Model construction and evaluation

A total of inactive (n=369,838) and active (n= 904) molecules from AID 743255 bioassay data was downloaded and using PowerMV 179 2D descriptors were created. Upon, post data processing using the feature selection method the total descriptors contributing to the generation of the predictive model came down to 45. The dataset was divided into two sets: (1) 90 % of the data as a training set, and (2) 10 % of the data

as an independent test set. After the implementation of the LDA algorithm to the preprocessed data set a predictive model was built and the statistical performance parameters of LDA algorithm are tabulated in Table I. An average accuracy of 96.76 % and 96.40 % was obtained for training and test data, respectively to screen active anti USP1 inhibitor was obtained upon 10 fold cross validation of AID 743255 dataset. Since accuracy alone is not sufficient to evaluate the efficiency of the model, therefore, another statistical parameter namely the AUC value was calculated from the ROC plot for both training and test set of data as shown in Fig. 2 and 3.

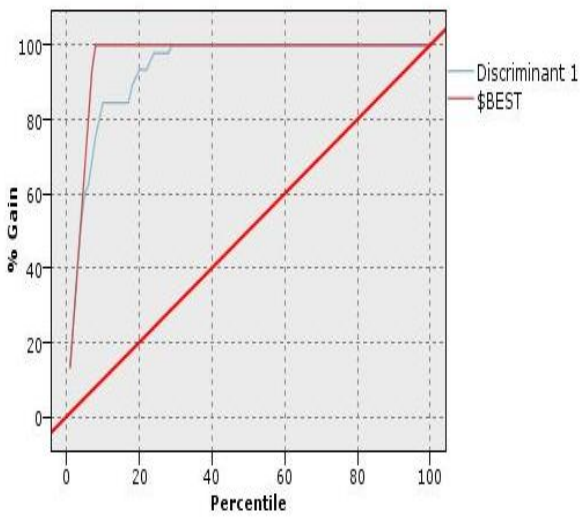

Fig. 3. Average prediction accuracy of LDA algorithm for 90 % training data set

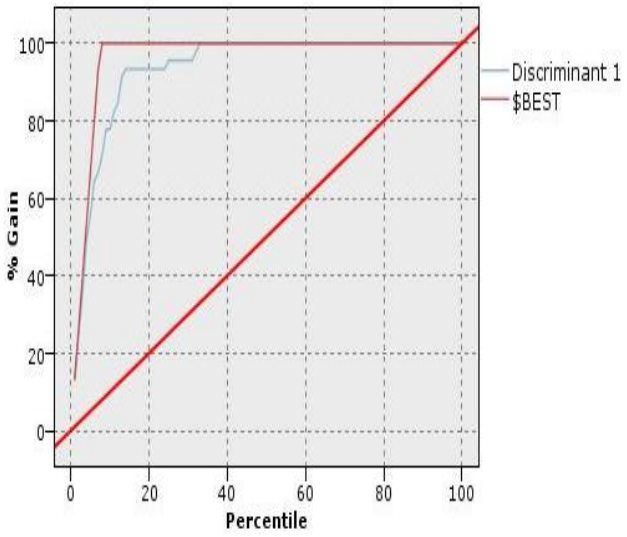

Fig. 4. Average prediction accuracy of LDA algorithm for 10 % testing data

The average value of AUC upon implementation of LDA algorithm to the training and independent test set of data was found to be 0.97 as shown in Table I.

TABLE. I. RESULTS ON THE AID 743255 DATASET USING LINEAR DISCRIMINANT ALGORITHM (LDA) ALGORITHM

Dataset Part	Accuracy		Error		Area under the curve	
	Training	Test	Training	Test	Training	Test
1	100	93.4	0	6.6	1	0.96
2	100	100.0	0	0	1	1.0
3	94.1	96.22	5.9	3.78	0.97	0.97
4	93.7	94.08	6.3	5.92	0.94	0.97
5	99.2	98.68	0.8	1.32	0.99	0.99
6	98.9	100.0	1.1	0	0.99	1.0
7	97.9	95.09	2.1	4.91	0.99	0.96
8	94.7	92.40	5.3	7.6	0.97	0.94
9	95.9	95.17	4.1	4.83	0.90	0.92
10	93.2	99.04	6.8	0.96	0.97	0.99
Average	**96.76**	**96.40**	**3.24**	**3.6**	**0.972**	**0.97**

As the AUC value of the predictive model is close to 1, therefore, we can propose that the chemoinformatics model generated using LDA classification algorithm will classify active anti USP1 inhibitor from any given dataset with high accuracy and specificity. All these statistics values were obtained by execution of the classification algorithm on the independent test set. The current predictive based on LDA classifier is more robust, efficient and accurate in predicting USP1 inhibitor molecule from AID 743255 dataset than the predictive model proposed by Wahi et.al [16].

The accuracy and AUC value of all the base classifier used by Wahi et al 2015, are lower than the present model which has a higher accuracy and AUC value as shown in Table II Therefore we say the present model is more robust and accurate in predicting active anti-cancer molecule having anti USP1 activity from a given dataset.

TABLE. II. COMPARATIVE PERFORMANCE EVALUATION OF CHEMOINFORMATICS MODELS

Algorithm	**Model performance evaluation parameters**		
	Accuracy	**Error**	**Area under the curve**
Random Forest	81.35	18.65	0.872
Naive Bayes	80.01	19.99	72.8
J48	80.1	19.9	78.3
SMO	80.21	19.79	78.7
Linear Discriminant Analysis (LDA)	96.76	3.24	0.97

IV. CONCLUSION

Targeting cancer by inhibiting USP1 is evolving as a promiscuous cancer therapy due to its specificity and efficacy when compared to the present-day anti-tumor remedies. The present drug discovery program involving experimental identification of a potent inhibitor of a target protein from huge chemical repositories is both a time taking and costly process.

The use of machine learning tools to analyze the huge data generated from high throughput screening (HTS) has paved the way to build a predictive chemoinformatics model for the screening of more anti-cancer molecule. In this regard, we have generated a computational predictive tool based on the properties and structure of known USP1 inhibitors from the high throughput screening experimental data. The present in silico predictive model can predict unknown inhibitors of the USP1/UAF1 complex with higher accuracy and reliability. The present chemoinformatics model generated using LDA algorithm has better accuracy to predict the anti USP1 activity of unknown compound when compared to random forest model proposed by Wahi et al in 2015. Our descriptor-based virtual screening computational predictive model will be of immense importance in prioritizing lead molecule against USP1/UAF1 complex and therefore fast-tracking the anti-USP1 drug discovery process. Moreover, the present chemical descriptor based predictive method can reduce the requisite for cost-intensive biological screening and encourage low-cost virtual screening on a larger scale to enhance the anti-cancer drug discovery process.

ACKNOWLEDGMENT

The faculty of computing and information technology at King Abdulaziz University, Saudi Arabia, supported this work. The authors would like also to thank Mr. Tabrej Khan for his assistance in procuring the required software for experimentation and the Faculty of Computing and Information Technology at Rabigh (FCITR) for providing the proper computational facility.

REFERENCES

[1] A.Y. Amerik and M.Hochstrasser, “Mechanism and function of deubiquitinating enzymes,” Biochim. Biophys. Acta., Vol. 1695, pp. 189–207, 2004.

[2] S.M. Nijman, M.P. Luna-Vargas, A.Velds, T.R. Brummelkamp, A.M. Dirac, T.K. Sixma, R. Bernards, “A genomic and functional inventory of deubiquitinating enzymes,” Cell, vol. 123, pp. 773–786, 2005b.

[3] J.M. Fraile, V. Quesada, D. Rodriguez, J.M. Freije, C. Lopez-Otin, “Deubiquitinases in cancer: new functions and therapeutic Options,” Oncogene, vol 31, pp. 2373–2388, 2012.

[4] S. Hussain, Y. Zhang and P.J. Galardy, “DUBs and cancer: the role of deubiquitinating enzymes as oncogenes, non-oncogenes and tumor suppressors,” Cell Cycle, vol 8, pp.1688–1697, 2009.

[5] J.J. Sacco, J.M. Coulson, M.J. Clague, S. Urbe, “Emerging roles of deubiquitinases in cancer-associated pathways,” IUBMB Life vol 62, pp. 140–157, 2010.

[6] D. Branzei and M. Foiani, “Regulation of DNA repair throughout the cell cycle,” Nat. Rev. Mol. Cell Biol., vol 9, pp. 297–308, 2008.

[7] R.D. Kennedy and A.D. D’Andrea, “DNA repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes,”J. Clin. Oncol., vol 24, pp. 3799–3808, 2006.

[8] T. Helleday, E. Petermann, C. Lundin, B. Hodgson, R.A. Sharma, “DNA repair pathways as targets for cancer therapy,” Nat. Rev. Cancer, vol 8, pp. 193–204, 2008.

[9] N.J. Curtin, “DNA repair dysregulation from cancer driver to the therapeutic target,” Nat. Rev. Cancer, vol 12, pp. 801–817, 2012.

[10] D. Hoeller and I. Dikic, “Targeting the ubiquitin system in cancer Therapy,” Nature vol 458, pp. 438–444, 2009.

[11] M.A. Cohn, P. Kowal, K.Yang, W. Haas, T.T. Huang, S.P. Gygi, A.D. D’Andrea, “A UAF1-containing multisubunit protein complex regulates the Fanconi anemia pathway,” Mol. Cell, vol 28, pp.786–797, 2007.

[12] M.A. Cohn, Y. Kee, W. Haas, S.P. Gygi, A.D. D’Andrea, “UAF1 is a subunit of multiple deubiquitinating enzyme complexes,” J. Biol. Chem., vol 284, pp. 5343–5351, 2009.

[13] M.A. Villamil, J. Chen, Q. Liang, Z. Zhuang, “A noncanonical cysteine protease USP1 is activated through active site modulation by USP1-associated factor 1,” Biochemistry, vol 51, pp. 2829–2839, 2012.

[14] Liang Q et al., “A selective USP1-UAF1 inhibitor links deubiquitination to DNA damage responses,” Nat. Chem. Biol. vol10, pp. 298–304, 2014.

[15] National Center for Biotechnology Information. PubChem BioAssay Database; AID=743255, https://pubchem.ncbi.nlm.nih.gov/bioassay/743255.

[16] D. Wahi, S. Jamal, S. Goyal, A. Singh, R. Jain, P. Rana and A. Grover, “Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents,”Syst. Synth. Biol., vol 9, pp. 33–43, 2015.

[17] M. Sud, “MayaChemTools,” 2010. http://www.mayachemtools.org/

[18] K. Liu, J. Feng, S.S.Young, “PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation,” J. Chem. Inf. Model, vol 45, pp. 515–522, 2005.

[19] M.L, Carlos, L. Belanche, and À. Nebot. "Feature selection algorithms: A survey and experimental evaluation." Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.

[20] A.Z. Dudek, T. Arodz, J. Galvez, “Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen,” vol 9, pp. 213–228, 2006.

[21] R.A, Caruana and D. Freitag. How Useful is Relevance? Technical report, Fall’94 AAAI Symposium on Relevance, New Orleans, 1994.

[22] S. Jamal and V. Scaria, “Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana,” BMC Bioinformatics, vol 14, pp. 329, 2013.

[23] S. Jamal and V. Scaria, “Predictive modeling of antimalarial molecules inhibiting apicoplast formation,” BMC Bioinformatics, vol 14, pp. 55, 2013.

[24] V. Periwal, J.K. Rajappan, A.U. Jaleel, V. Scaria, “Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets,” BMC Res. Notes, vol 4, pp. 504, 2011.

[25] V. Periwal, S. Kishtapuram, V. Scaria, “Computational models for in vitro anti-tubercular activity of molecules based on high throughput chemical biology screening datasets,” BMC Pharmacol. vol 12, pp.1, 2012.

[26] D. Huang, Y. Quan, M. He, and B. Zhou, “Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data,”J. Exp. Clin. Cancer Research, vol 28, pp.149, 2009.

[27] M. Mathea, W. Klingspohn, and K. Baumann, “Chemoinformatic Classification Methods and their Applicability Domain,” Mol. Inf., vol 35, pp. 160 – 180, 2016.

[28] Fisher, “The use of multiple measurements in taxonomic problems’. Annals of Eugenics,” vol 7, pp. 179–188, 1936.

[29] T. Li, S. Zhu, M. Ogihara. "Using discriminant analysis for multi-class classification: an experimental investigation." Knowl Inf Syst. Vol 10.4, pp. 453-472, 2006.

[30] M. Abdullahand and A.I. Ghoson, "Decision tree induction and clustering techniques in sas enterprise miner, spss clementine, and IBM intelligent minera comprehensive analysis," International Journal of Management and Information Systems, vol 14, pp. 57-70. 2010.