ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Machine learning-driven prediction of EBNA1 inhibitors against Epstein–Barr virus in nasopharyngeal carcinoma

Lavinia Clarisa Wicklem[1], Siaw San Hwang[1], Bee Theng Lau[1], Mrinal Bhave[2], and Xavier Wezen Chee[1]*

[1]Science Programme, School of Engineering and Science, Swinburne University of Technology (Sarawak Campus), Kuching, Sarawak, Malaysia

[2]Department of Chemistry and Biotechnology, School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Victoria, Australia

Abstract

Nasopharyngeal carcinoma (NPC), particularly prevalent in regions such as Malaysia, is a significant health concern often linked to Epstein-Barr virus (EBV) infection. The EBV nuclear antigen 1 (EBNA1), crucial for EBV survival and NPC tumorigenicity, has emerged as a potential therapeutic target for EBV-positive NPC. In this study, we utilized quantitative structure-activity relationship (QSAR) models to predict potential inhibitors of EBNA1. These models were developed based on the molecular fingerprints of known EBNA1 inhibitors, using both classification and regression approaches. Our QSAR classification models demonstrated consistently high precision, recall, F1 score, and accuracy scores across the training set. The top-performing models, constructed using logistic regression algorithms, achieved perfect precision scores of 1.000 in the test set evaluation. These models' recall, F1 score, and accuracy scores were 0.571, 0.727, and 0.667, respectively. On the other hand, the best-performing model among the regression models was built using the sequential minimal optimization regression algorithm, achieving a correlation coefficient of 0.703. The mean absolute error and root mean square error of this QSAR regression model were 0.173 and 0.217, respectively, whereas the relative absolute error was 0.689. We screened the enamine advanced compound library using this regression model to predict compounds with potential EBNA1 inhibitory effects. This led to the identification of the top 10 compounds with the most promising predicted EBNA1 inhibitory properties.

Keywords: Epstein-Barr virus nuclear antigen 1; Nasopharyngeal carcinoma; Quantitative structure-activity relationship; Inhibitor; Machine learning; Compound screening

***Corresponding author:**
Xavier Wezen Chee
(xchee@swinburne.edu.my)

Citation: Wicklem LC, Hwang SS, Lau BT, Bhave M, Chee XW. Machine learning-driven prediction of EBNA1 inhibitors against Epstein–Barr virus in nasopharyngeal carcinoma. *Artif Intell Health*. 2025;2(1):93-104. doi: 10.36922/aih.4375

Received: July 30, 2024

Revised: September 10, 2024

Accepted: September 23, 2024

Published Online: November 8, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The drug discovery process involves several stages, starting with the identification of disease targets and the search for small molecules that can modulate these targets. This often involves testing thousands to millions of compounds in various assays, with only a few progressing to animal testing and pre-clinical studies.[1] Conclusively, developing new and effective drugs is tedious, requiring millions of dollars and spanning over a decade.[2]

While high-throughput screening can identify active compounds at low concentrations, it often produces a low hit rate or high false positives.[3] This leads to a significant discrepancy between the number of hits identified and the number of viable lead compounds, which leads to wastage.

One way to negate this problem is using quantitative structure-activity relationship (QSAR) in drug discovery. QSAR is a ligand-based drug design method that uses mathematical models to correlate the chemical features of inhibitors to their bioactivity.[4] QSAR models streamline drug discovery by predicting compound activity based on their structure and properties, allowing researchers to prioritize promising candidates. This targeted approach reduces the need for extensive testing, saving time, resources, and materials while accelerating the research process. It optimizes resource allocation and promotes sustainable research practices by focusing efforts on compounds with a higher likelihood of success.[5] Another advantage is that QSAR aids in designing active molecules in a "greener" way by reducing the need for extensive experimental synthesis and testing on animals. A study examined the toxicity of various ionic liquids (ILs), which have the potential to harm aquatic life. Their study utilized advanced QSAR techniques to develop reliable models for predicting IL toxicity without animal testing. Their QSAR models demonstrated high predictive accuracy, with classification models achieving over 86% accuracy and regression models showing a correlation (R^2) >0.90 in the test data. These high-performance models provided strong predictions and pinpointed the structural elements of ILs contributing to their cytotoxicity. These QSAR models offer valuable tools for designing safer and environmentally friendly ILs.[6] Notably, QSAR-based virtual screening has emerged as a pivotal approach in contemporary scientific investigations, facilitating the identification of potential drug candidates. QSAR has been used to design chalcone derivatives that outperform standard tuberculosis drugs, identify potent neuraminidase inhibitors for influenza A, identify potent inhibitors for 5-HT_{1A} receptors for mood and anxiety disorders, and identify potential antimalarial activity in compounds that have low toxicity towards the mammalian cell.[7-10] QSAR was also used to identify critical structural features enhancing the inhibitory effects of compounds against liver carcinoma cells in tumor-targeting drug studies.[11] In antipsychotic/antidepressant studies, QSAR models have aided in predicting the activities of natural compounds against specific receptors, offering potential alternatives to synthetic drugs.[12] QSAR methodologies were also used to clarify physicochemical factors influencing the activity and cytotoxicity of compounds against human immunodeficiency viruses and influenza viruses in antiviral drug studies.[13,14]

Nasopharyngeal carcinoma (NPC) is strongly associated with Epstein-Barr virus (EBV). NPC typically affects individuals in their mid-40s and is more prevalent in men. It consistently exhibits EBV positivity, regardless of geographic location. Annually, NPC accounts for approximately 90,000 cases and 50,000 deaths recorded globally.[15] Its distribution is unique, with Asian countries representing around 80% of documented cases and mortality rates. In Malaysia, NPC ranks as the fourth most common cancer among males.[16] Among the Bidayuh community, part of Malaysia's indigenous population, the risk of NPC is notably elevated, with men and women experiencing a 2.3-fold and 1.9-fold increase, respectively, compared to other populations during the same period.[17] NPC poses a significant health concern, among which EBV latent infection stands out as a prominent contributor.

EBV is a virus capable of infecting epithelial and B cells, facilitating its persistence within the host and transmission among humans. A critical protein in maintaining viral stability and promoting viral gene expression is called the EBV nuclear antigen 1 (EBNA1). EBNA1 interacts with the oriP region of the EBV genome, forming dimers and complexes crucial for DNA looping and maintaining genome stability.[18,19] In addition, EBNA1 recruits cellular proteins to facilitate DNA replication.[20] EBNA1 binds to the Family of Repeats (FR) element during cell division, tethering EBV episomes to cellular chromosomes for proper segregation.[21-23] EBNA1 also activates EBV gene transcription by interacting with the FR element, with specific regions within EBNA1 being crucial for this function.[24] Moreover, EBNA1 affects several cellular signaling pathways in cell transformation and growth. It amplifies STAT1 signaling, enhances interferon responsiveness, and inhibits the transforming growth factor beta and nuclear factor kappa B pathways, ultimately promoting viral persistence and oncogenesis.[25,26] EBNA1 also disrupts promyelocytic nuclear bodies, impairing DNA repair, p53 activation, and apoptosis in response to DNA damage.[27] This disruption is mediated by interactions with cellular proteins ubiquitin-specific-processing protease 7 (USP7) and casein kinase 2, leading to promyelocytic leukemia protein degradation.[27-29] EBNA1 interacts with USP7, stabilizing its binding and preventing p53 stabilization protease.[30] Consequently, cells expressing EBNA1 exhibit reduced p53 accumulation upon DNA damage, facilitating cell survival and potentially contributing to tumorigenesis.[27] EBNA1 expression also correlates with increased oxidative stress, characterized by elevated reactive oxygen species (ROS) levels and DNA damage. This phenomenon, mediated by the upregulation of ROS-generating enzyme NADPH oxidase 2 (NOX2), may promote genomic instability and tumor development.[31,32]

Due to the involvement of EBNA1 in EBV's persistence and oncogenesis, we decided to deploy QSAR modeling to identify inhibitors targeting EBNA1. At present, QSAR applications in search of EBNA1 inhibitors remain unexplored in the current scientific literature. To bridge this gap, our research aims to identify potential compounds with inhibitory activities against EBNA1 using our QSAR models.

2. Data and methods

2.1. Dataset preparation

We developed the QSAR models using the AID2381 dataset obtained from a study by Gianti *et al.*[33] into molecular descriptors and fingerprints. All the compounds in the dataset demonstrated inhibitory activity toward EBNA1 through *in vitro* studies. The compounds in the database were experimentally evaluated using fluorescence polarization assay and were shown to inhibit EBNA1 selectively. First, we split the dataset into a training set and an external test set with a ratio of approximately 4:1. This yields a training set with 34 compounds and a test set with nine compounds. The compounds from these two datasets were then featured with chemical fingerprints using the PaDEL-Descriptor package. In total, 1024 chemical fingerprints were generated for each chemical compound in both datasets. After conversion into chemical fingerprints, we cleaned the dataset by removing empty rows and columns. In addition, we extracted the bioactivity of the ligands in pIC_{50} format.

2.2. Attribute selection

We constructed the QSAR models using the Waikato Environment of Knowledge Analysis (WEKA) package.[34] WEKA is a software consisting of an extensive collection of machine learning algorithms for data mining and exploration.[35] Before model construction, we performed attribute selection to identify the most relevant features for the model construction.[36] There are two parts to selecting the attributes: Attribute evaluation and search method. The attribute evaluation assesses each attribute related to the output variable within the dataset. We applied two methods of attribute evaluation: CfsSubsetEval (CFS) and ClassifierSubsetEval (CSE).

2.2.1. CFS

This method evaluates the worth of a subset of attributes by considering each feature's predictive ability and the degree of redundancy between them. Subsets of features highly correlated with the class while having low intercorrelation are preferred.[37] To select attributes, the attribute evaluator will employ a search method. The search method systematically explores various combinations of attributes within the dataset, aiming to identify a selection of preferred features. We used two search methods for CFS: Best first (BF) and greedy stepwise (GS). The BF method searches the attribute space by greedy hill climbing augmented with a backtracking facility, while the GS method performs a greedy forward or backward search through the space of attribute subsets.[38,39]

2.2.2. CSE

This method uses an algorithm to estimate the "merit" of attributes.[37] We used several algorithms for CSE to select the top attributes. For classification modeling, we employed algorithms Naïve Bayes (NB), instance-based learner (IBK), J48 Decision Tree (J48), random forest (RF), and logistic regression (LR). For regression modeling, we used linear regression (LRE), simple linear regression (SLR), sequential minimal optimization (SMO) regression, IBK, and RF algorithms. We also employed search methods BF and GS for CSE. For better visualization, we show the attribute selection process in this study (Figure 1).

2.3. Classification QSAR model

After the attribute selection process, we built the classification models using the NB, IBK, J48, RF, and LR algorithms.

2.3.1. Evaluation metrics for classification

The performance of the classification model was evaluated using standard metrics, including precision, recall, F1 score, and accuracy. Precision is a metric that evaluates the accuracy of correct predictions. It is calculated by dividing the number of accurate positive predictions by the total number of positive predictions.[40]

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(I)}$$

where TP is true positive, and FP is false positive.

The recall metric measures the number of actual observations predicted correctly. It is determined by dividing the number of correct positive predictions by the total number of actual positive instances.[40]

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{(II)}$$

where TP is true positive, and FN is false negative.

F1 score is a metric that calculates the harmonic mean between precision and recall. The formula of F1 score, which provides a balanced measure of a model's performance, is given as follows:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} \times \text{recall}} \quad \text{(III)}$$

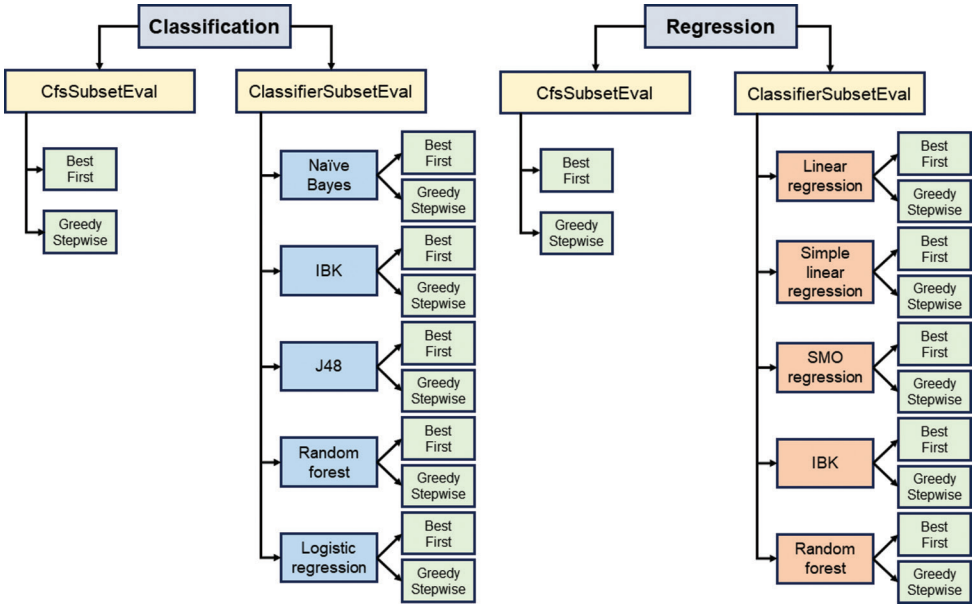

Figure 1. Quantitative structure-activity relationship model process
Abbreviations: IBK: Instance-based learner; SMO: Sequential minimal optimization.

Accuracy is the sum of two accurate predictions divided by the total number of data sets. It measures how often the classifier makes the correct prediction. It is the ratio between the number of correct predictions and the total number of forecasts.[40] We can calculate accuracy using the formula below.

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

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

These four evaluation classification metrics can provide a comprehensive understanding of the performance of the classification QSAR models.

2.4. Regression QSAR model

After the attribute selection process, we built the QSAR regression models using LRE, SLR, SMO regression, IBK, and RF algorithms.

2.4.1. Evaluation metrics for regression

We assessed the regression QSAR models' performance using correlation coefficient (r), mean absolute error (MAE), root mean squared error (RMSE), and relative absolute error (RAE) scores. The R score is a statistical measure of the strength of a linear relationship between two variables. The value of r ranges from −1 to 1. A negative score indicates an inverse correlation between the variables, whereas a positive score means the variables have a positive correlation.[41] Meanwhile, an r value close to 0 indicates a very weak or no linear correlation between the variables.[41] r is calculated as below.]

$$r = \frac{\sum\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 * \sum\left(y_i - \bar{y}\right)^2}} \quad \text{(V)}$$

where: x_i: each individual x value; $\bar{x}$: mean of all x values; y_i: each individual y value; $\bar{y}$: mean of all y values

MAE score is calculated as the average of the absolute error values between the observed and predicted values. The score ranges from 1 being perfect to 0 being wrong.[42] MAE is calculated as below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \check{y}_i\right| \quad \text{(VI)}$$

wheren represents the number of predictions; y_i represents the observed values; and $\hat{y}_i$ represents the predicted values.

RMSE is the squared root of the mean of all the errors, which describes the prediction magnitude error.[43] The scores range from 1 to 0, with lower scores preferred. RAE is determined by dividing the sum of absolute errors by the absolute difference between the mean and the actual value. The equation for RMSE is given in the following.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad \text{(VII)}$$

wheren represents the number of predictions; y_i represents the observed values; and $\hat{y}_i$ represents the predicted values.

RAE serves as a measure to assess the performance of a predictive model and is represented as a ratio. Lower RAE scores indicate a more effective mode.[44] The equation for calculating RAE is as follows.

$$RAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|} \quad \text{(VIII)}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{(IX)}$$

where n represents the number of observations; y_i represents the observed value; and $\bar{y}$ represents the average of observed values.

These four assessment regression metrics offer a thorough perspective on the performance of regression QSAR models.

2.5. Model deployment

After constructing the QSAR models, we validated all our models using the external test set. The chosen model was then deployed on the enamine advanced compound library. The enamine advanced compound library was similarly featurized with chemical fingerprints using the PaDEL-Descriptor package.

3. Results

3.1. Classification QSAR models

Our study yielded the following results for classification-based machine learning models (Table 1). CFS-LR-BF and CFS-LR-GS models exhibited precision scores of 1.000, recall scores of 0.952, F1 scores of 0.976, and accuracy scores of 0.976. In addition, the CFS-NB-BF and CFS-NB-GS models had precision, recall, and F1 scores all at 0.952 and accuracy scores at 0.953. The CSE-J48-LR-BF model achieved a precision score of 0.955, a recall score of 1.000, an F1 score of 0.977, and an accuracy score of 0.976. Meanwhile, the CSE-J48-IBK-BF model demonstrated a precision score of 0.952, a recall score of 0.952, an F1 score of 0.952, and an accuracy score of 0.953. We visualized the performance of these models using a confusion matrix (Figure 2).

We evaluated our models using an external test set comprising eight compounds (Table 2). The CFS-LR-BF and CFS-LR-GS QSAR classification models demonstrated precision scores of 1.000 and recall scores of 0.571. Both models achieved F1 scores of 0.727 and accuracy scores of 0.667. The CFS-NB-BF and CFS-NB-GS models exhibited precision scores of 1.000 and recall scores of 0.429. Both models achieved F1 scores of 0.600 and accuracy scores of 0.556. Finally, the CSE-J48-LR-BF and CSE-J48-IBK-BF models demonstrated precision scores of 1.000, with recall scores of 0.429. Both models achieved F1 scores of 0.600 and accuracy scores of 0.556. We also presented the results of the test set evaluation using a series of confusion matrices (Figure 3). These visual representations show the models' performance in classifying active and inactive compounds.

3.2. Regression QSAR models

For regression-based models, we obtained the following results. For the training set of CSE-LRE-BF-SMO and CSE-LRE-GS-SMO, both models achieved R scores of 0.992. Both models had MAE values of 0.029 and RMSE values of 0.037. The RAE values for both models were 0.118. For the training set of the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE QSAR regression models, both models achieved R scores of 0.999. Both models had MAE values of 0.004 and RMSE values of 0.005. The RAE values for both models were 0.014. Regarding the training set results for the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO QSAR regression models, we observed that both models achieved R scores of 0.999. Both models achieved MAE values of 0.008 and RMSE values of 0.010. The RAE values for both models were 0.032. We plotted the graphs of experimental pIC_{50} versus predicted pIC_{50} of the compounds in the training set (Figure 4). Consecutively, we evaluated the models on a test set to determine the predictive power of each model.

Table 1. Score for evaluation metric for the training set

	CFS-LR-BF	CFS-LR-GS	CFS-NB-BF	CFS-NB-GS	CSE-J48-LR-BF	CSE-J48-IBK-BF
Precision	1.000	1.000	0.952	0.952	0.955	0.952
Recall	0.952	0.952	0.952	0.952	1.000	0.952
F1 score	0.976	0.976	0.952	0.952	0.977	0.952
Accuracy	0.976	0.976	0.953	0.953	0.976	0.953

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

Figure 2. Confusion matrix for the training set results of (A) CFS-LR-BF, (B) CFS-LR-GF, (C) CSE-J48-BF-LR, (D) CFS-NB-BF, (E) CFS-NB-GS, and (F) CSE-J48-BF-IBK

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

Table 2. Score for evaluation metric for the test set

	CFS-LR-BF	CFS-LR-GS	CFS-NB-BF	CFS-NB-GS	CSE-J48-LR-BF	CSE-J48-IBK-BF
Precision	1.000	1.000	1.000	1.000	1.000	1.000
Recall	0.571	0.571	0.429	0.429	0.429	0.429
F1 score	0.727	0.727	0.600	0.600	0.600	0.600
Accuracy	0.667	0.667	0.556	0.556	0.556	0.556

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

For our external test set results, we observed that the CSE-LRE-BF-SMO and CSE-LRE-GS-SMO achieved R scores of 0.703 and 0.705, respectively. The MAE and RMSE values for both models were 0.173 and 0.217, respectively. Meanwhile, the RAE values for both models were 0.688 and 0.686, respectively. Both the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE QSAR regression models achieved an R score of 0.703 in the test set. The MAE and RMSE values were 0.173 and 0.217, respectively. The RAE values for both models were 0.689. Moving on to the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO QSAR regression models, both models achieved an R score of 0.703 in the test set. The MAE values for both models were 0.173 whereas the RMSE values for both models were 0.217. The RAE values for both models were 0.689. The outcomes of the test set evaluation are depicted through a table summarizing the different evaluation metrics (Table 3) and plots of actual pIC_{50} versus predicted pIC_{50} (Figure 5).

3.3. Deployment of model

Given that our target variable is the pIC_{50} of compounds, we decided to employ a modeling approach that provides numerical outcomes, namely the regression algorithm. Therefore, we chose to deploy the CSE-SMO-BF-LRE model on the enamine advanced library to predict their inhibitory activities against EBNA1. After the enamine advanced library compounds were featured with chemical fingerprints, we predicted their pIC_{50} against EBNA1 using the chosen regression model. The structures of the top 10 compounds are shown in Figure 6. Future work would involve purchasing these ten compounds for experimental validation.

4. Discussion

4.1. Classification QSAR models

We assessed our classification QSAR models' performance using four key metrics: Precision, recall, F1 score, and

Figure 3. Confusion matrix for the test set results of (A) CFS-LR-BF, (B) CFS-LR-GF, (C) CSE-J48-BF-LR, (D) CFS-NB-BF, (E) CFS-NB-GS, and (F) CSE-J48-BF-IBK

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

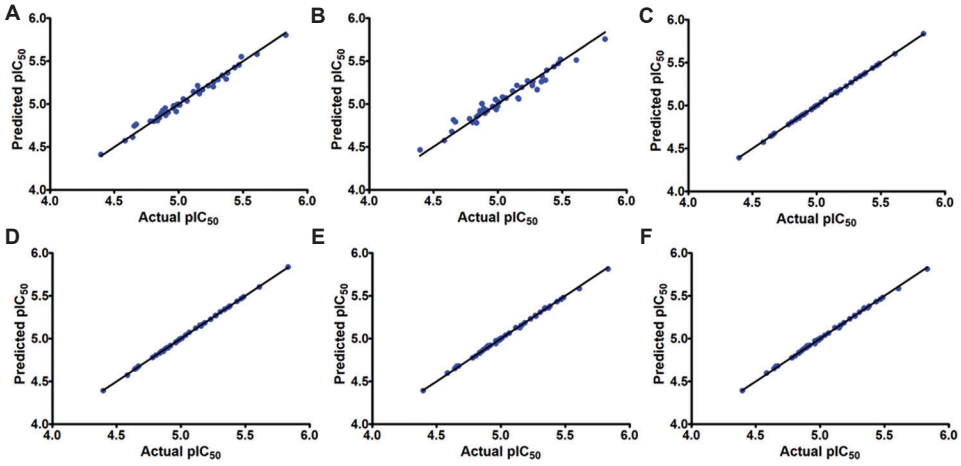

Figure 4. Graphical representation for the training set result for (A) CSE-LRE-BF-SMO, (B) CSE-LRE-GS-SMO, (C) CSE-SMO-BF-LRE, (D) CSE-SMO-GS-LRE, (E) CSE-SMO-BF-SMO, and (F) CSE-SMO-GS-SMO

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; SMO: Sequential minimal optimization.

Table 3. Score for evaluation metric for the test set

	CSE-LRE-BF-SMO	CSE-LRE-GS-SMO	CSE-SMO-BF-LRE	CSE-SMO-GS-LRE	CSE-SMO-BF-SMO	CSE-SMO-GS-SMO
R	0.703	0.705	0.703	0.703	0.703	0.703
MAE	0.173	0.172	0.173	0.173	0.173	0.173
RMSE	0.217	0.217	0.217	0.217	0.217	0.217
RAE	0.688	0.686	0.689	0.689	0.689	0.689

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; MAE: Mean absolute error; R: Correlation coefficient; RAE: Relative absolute error; RMSE: Root mean squared error; SMO: Sequential minimal optimization.

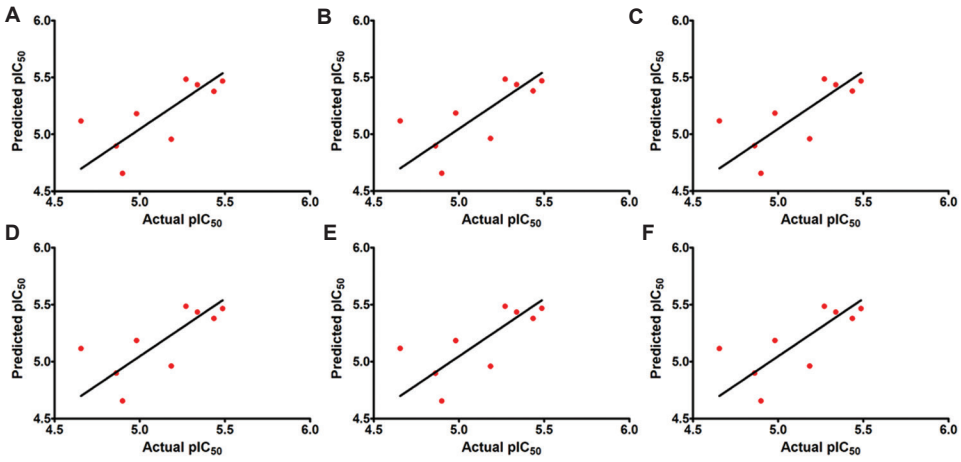

Figure 5. Graphical representation for the test set result for (A) CSE-LRE-BF-SMO, (B) CSE-LRE-GS-SMO, (C) CSE-SMO-BF-LRE, (D) CSE-SMO-GS-LRE, (E) CSE-SMO-BF-SMO, and (F) CSE-SMO-GS-SMO
Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; SMO: Sequential minimal optimization.

Table 4. Score for evaluation metric for the training set

	CSE-LRE-BF-SMO	CSE-LRE-GS-SMO	CSE-SMO-BF-LRE	CSE-SMO-GS-LRE	CSE-SMO-BF-SMO	CSE-SMO-GS-SMO
R	0.992	0.992	0.999	0.999	0.999	0.999
MAE	0.029	0.029	0.004	0.004	0.008	0.008
RMSE	0.037	0.037	0.005	0.005	0.010	0.010
RAE	0.118	0.118	0.014	0.014	0.032	0.032

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; MAE: Mean absolute error; R: Correlation coefficient; RAE: Relative absolute error; RMSE: Root mean squared error; SMO: Sequential minimal optimization.

accuracy. Our results highlighted two top-performing classification models, CFS-LR-BF and CFS-LR-GS. Both models exhibited high precision, recall, F1, and accuracy scores. In addition, the rest of the classification models also demonstrated strong performance (Figure 2). Our results showed that all six models accurately and successfully classified active and inactive compounds in the training set. During the external test set evaluation (Table 2), the CFS-LR-BF and CFS-LR-GS QSAR classification models demonstrated perfect precision scores of 1.000, indicating their precision in classifying a compound as active. However, their recall scores were moderate at 0.571, suggesting some active compounds might have been missed. Both models achieved F1 scores of 0.727 and accuracy scores of 0.667, indicating a balanced performance. The CFS-NB-BF and CFS-NB-GS models also exhibited perfect precision scores of 1.000, but their recall scores were lower at 0.429. Both models achieved consistent F1 scores of 0.600 and accuracy scores of 0.556. Finally, the CSE-J48-LR-BF and CSE-J48-IBK-BF models demonstrated perfect precision scores of 1.000, with moderately low recall scores of 0.429. Both models achieved consistent F1 scores of 0.600 and accuracy scores of 0.556. The consistently high precision scores across all models indicate their ability to identify active compounds correctly. However, the variability in recall scores suggests differences in their abilities to capture all true positive instances. While the models excel in minimizing false positive predictions, they may have limitations in identifying all active compounds in the dataset. Considering the scores of all models, we suggest that CFS-LR-BF and CFS-LR-GS are the top QSAR models for classification tasks.

4.2. Regression QSAR models

The performance of our regression-based QSAR models was evaluated using several key metrics: The correlation coefficient (R), MAE, RMSE, and RAE (Table 4). Based on the training set scores for the QSAR regression models, all models achieved high R scores with low MAE and RMSE values. Consequently, all the regression QSAR models demonstrated excellent predictive performance, with high correlation, low error rates, and minimal relative error in the training set. However, a good model cannot be determined solely by good scores on the training set. Therefore, we also evaluated the models on a test set to determine the predictive power of each model. Based on our external test set results, we observed that the CSE-LRE-BF-SMO and CSE-LRE-GS-SMO regression QSAR models

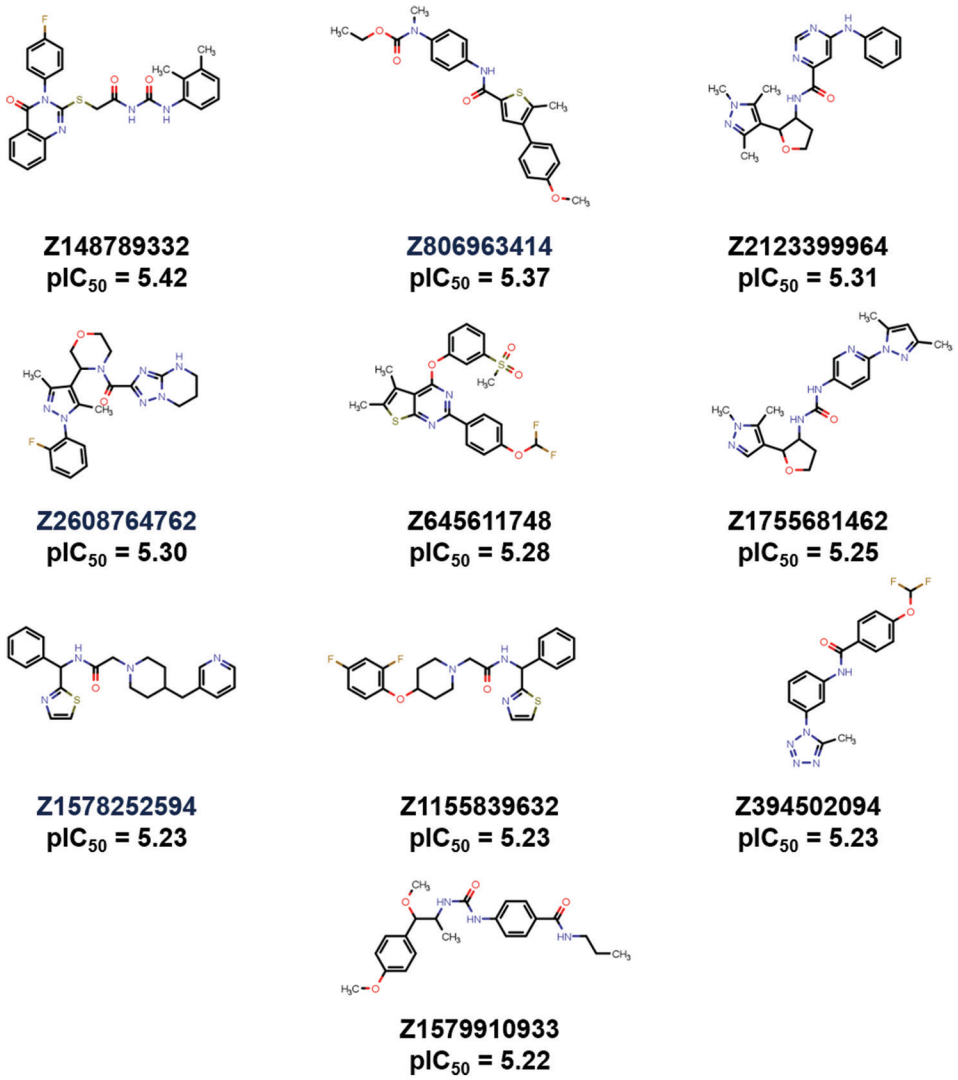

Figure 6. Two-dimensional chemical structures of chosen compounds generated using MarvinSketch 23.12

performed the best. Both models achieved R scores of 0.703 and 0.705, respectively. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. These error values suggest that the models' predictions deviate from the actual values by a small amount.[45] Meanwhile, the RAE values for both models were moderate, with values of 0.688 and 0.686, respectively. The RAE scores suggest that the models' predictions deviate from the actual values by a moderate percentage relative to the scale of the target variable. For the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE regression QSAR models, both models achieved an R score of 0.703 in the test set. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. The RAE values for both models were also moderate, at 0.689. Moving on to the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO regression QSAR models, both models achieved an R score of 0.703 in the test set. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. The RAE values for both models were also moderate, at 0.689. The outcomes of the test set evaluation are depicted through a table summarizing the different evaluation metrics (Table 3) and plots of actual pIC_{50} versus predicted pIC_{50} (Figure 5).

5. Conclusion

This study highlights the potential of QSAR modeling in identifying candidate compounds for inhibiting EBNA1, a key target in addressing EBV-associated diseases such as NPC. Our findings demonstrated that QSAR classification models, particularly CFS-LR-BF and CFS-LR-GS, exhibit strong precision, albeit with moderate recall. This suggests

their effectiveness in identifying active compounds while minimizing false positives. Despite the moderate recall, their balanced F1 scores and moderate accuracy indicate good performance. Similarly, the CSE-SMO-BF-LRE QSAR model captured the relationship between compound bioactivity and chemical fingerprints. Using QSAR for our drug screening process optimized resource allocation and reduced the need for extensive experimental synthesis, aligning with sustainable research practices. Furthermore, our QSAR-based screening of the Enamine Advanced compound library predicted the top 10 compounds with potential inhibitory effects against EBNA1. Further experimental validation of these predicted inhibitors is needed to confirm their efficacy and safety, paving the way for potential therapeutic interventions against EBV-positive NPC.

Acknowledgments

L.C.W. was supported by the Swinburne Sarawak Postgraduate Scholarship.

Funding

This work was supported by the MAKNA Cancer Research Award 2021 given to Xavier Wezen Chee.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Xavier Wezen Chee
Formal analysis: Lavinia Clarisa Wicklem, Bee Theng Lau, Xavier Wezen Chee
Investigation: Lavinia Clarisa Wicklem, Xavier Wezen Chee
Methodology: Siaw San Hwang, Bee Theng Lau, Mrinal Bhave, Xavier Wezen Chee
Writing – original draft: Lavinia Clarisa Wicklem
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be made available from the corresponding authors upon reasonable request.

Further disclosure

The authors would like to disclose that part of the findings included in the manuscript have been presented on (date: 3 August 2023) at the Swinburne Sarawak Postgraduate Research Conference 2023, (location: Swinburne University of Technology Sarawak).

References

1. Singh N, Vayer P, Tanwar S, Poyet JL, Tsaioun K, Villoutreix BO. Drug discovery and development: Introduction to the general public and patient groups. *Front Drug Discov.* 2023;3:1201419.

   doi: 10.3389/fddsv.2023.1201419

2. Sun J, Warden AR, Ding X. Recent advances in microfluidics for drug screening. *Biomicrofluidics*. 2019;13(6):061503.

   doi: 10.1063/1.5121200

3. Thorne N, Auld DS, Inglese J. Apparent activity in high-throughput screening: Origins of compound-dependent assay interference. *Curr Opin Chem Biol.* 2010;14(3):315-324.

   doi: 10.1016/j.cbpa.2010.03.020

4. Hansch C, Fujita T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc.* 1964;86(8):1616-1626.

   doi: 10.1021/ja01062a035

5. Chatterjee A. 27 - Computational methods and tools for sustainable and green approaches in drug discovery. In: Banik BK, editor. *Green Approaches in Medicinal Chemistry for Sustainable Drug Design.* Amsterdam: Elsevier; 2020. p. 965-988.

   doi: 10.1016/B978-0-12-817592-7.00027-7

6. Gupta S, Basant N, Singh KP. Nonlinear QSAR modeling for predicting cytotoxicity of ionic liquids in leukemia rat cell line: An aid to green chemicals designing. *Environ Sci Pollut Res.* 2015;22:12699-12710.

   doi: 10.1007/s11356-015-4526-3

7. Gomes MN, Braga RC, Grzelak EM, *et al*. QSAR-driven design, synthesis and discovery of potent chalcone derivatives with antitubercular activity. *Eur J Med Chem.* 2017;137:126-138.

   doi: 10.1016/j.ejmech.2017.05.026

8. Lian W, Fang J, Li C, Pang X, Liu AL, Du GH. Discovery of influenza a virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models. *Mol Divers.* 2016;20(2):439-451.

   doi: 10.1007/s11030-015-9641-z

9. Luo M, Wang XS, Roth BL, Golbraikh A, Tropsha A. Application of quantitative structure-activity relationship models of 5-HT1A receptor binding to virtual screening identifies novel and potent 5-HT1A ligands. *J Chem Inf Model.* 2014;54(2):634-647.

   doi: 10.1021/ci400460q

10. Zhang L, Fourches D, Sedykh A, *et al*. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J Chem Inf Model.* 2013;53(2):475-492.

doi: 10.1021/ci300421n

11. Kamano Y, Yamashita A, Nogawa T, *et al*. QSAR evaluation of the Ch'an Su and related bufadienolides against the colchicine-resistant primary liver carcinoma cell line PLC/PRF/5. *J Med Chem*. 2002;45(25):5440-5447.

doi: 10.1021/jm0202066

12. Avram S, Stan MS, Udrea AM, Buiu C, Boboc AA, Mernea M. 3D-ALMOND-QSAR models to predict the antidepressant effect of some natural compounds. *Pharmaceutics*. 2021;13(9):1449.

doi: 10.3390/pharmaceutics13091449

13. Ravichandran V, Jain A, Mourya V, Agrawal RK. Prediction of anti-HIV activity and cytotoxicity of pyrimidinyl and triazinyl amines: A QSAR study. *Chem Pap.* 2008;62:596-602.

doi: 10.2478/s11696-008-0072-5

14. Yuan H, Parrill AL. QSAR studies of HIV-1 integrase inhibition. *Bioorg Med Chem.* 2002;10(12):4169-4183.

doi: 10.1016/s0968-0896(02)00332-2

15. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.

doi: 10.3322/caac.21492

16. Azizah AM, Hashimah B, Nirmal K, *et al*. *Malaysia National Cancer Registry Report* (*MNCR*). Putrajaya, Malaysia: National Cancer Institute, Ministry of Health; 2019.

17. Devi BCR, Pisani P, Tang TS, Parkin DM. High incidence of nasopharyngeal carcinoma in native people of Sarawak, Borneo Island. *Cancer Epidemiol Biomarkers Prev.* 2004;13(3):482-486.

18. Yates J, Warren N, Reisman D, Sugden B. A cis-acting element from the Epstein-Barr viral genome that permits stable replication of recombinant plasmids in latently infected cells. *Proc Natl Acad Sci U S A*. 1984;81(12):3806-3810.

doi: 10.1073/pnas.81.12.3806

19. Frappier L, O'Donnell M. Epstein-barr nuclear antigen 1 mediates a DNA loop within the latent replication origin of Epstein-Barr virus. *Proc Natl Acad Sci U S A*. 1991;88(23):10875-10879.

doi: 10.1073/pnas.88.23.10875

20. Chaudhuri B, Xu H, Todorov I, Dutta A, Yates JL. Human DNA replication initiation factors, ORC and MCM, associate with oriP of Epstein-Barr virus. *Proc Natl Acad Sci U S A*. 2001;98(18):10085-10089.

doi: 10.1073/pnas.181347998

21. Harris A, Young BD, Griffin BE. Random association of Epstein-Barr virus genomes with host cell metaphase chromosomes in Burkitt's lymphoma-derived cell lines. *J Virol.* 1985;56(1):328-332.

doi: 10.1128/JVI.56.1.328-332.1985

22. Petti L, Sample C, Kieff E. Subnuclear localization and phosphorylation of Epstein-Barr virus latent infection nuclear proteins. *Virology*. 1990;176(2):563-574.

doi: 10.1016/0042-6822(90)90027-o

23. Lee MA, Diamond ME, Yates JL. Genetic evidence that EBNA-1 is needed for efficient, stable latent infection by Epstein-Barr virus. *J Virol.* 1999;73(4):2974-2982.

doi: 10.1128/jvi.73.4.2974-2982.1999

24. Lupton S, Levine AJ. Mapping genetic elements of Epstein-Barr virus that facilitate extrachromosomal persistence of Epstein-Barr virus-derived plasmids in human cells. *Mol Cell Biol*. 1985;5:2533-2542.

doi: 10.1128/mcb.5.10.2533-2542.1985

25. Wood VHJ, O'Neil JD, Wei W, Stewart SE, Dawson CW, Young LS. Epstein-Barr virus-encoded EBNA1 regulates cellular gene transcription and modulates the STAT1 and TGFbeta signaling pathways. *Oncogene.* 2007;26(28):4135-4147.

doi: 10.1038/sj.onc.1210496

26. Valentine R, Dawson CW, Hu C, *et al*. Epstein-Barr virus-encoded EBNA1 inhibits the canonical NF-κB pathway in carcinoma cells by inhibiting IKK phosphorylation. *Mol Cancer.* 2010;9:1.

doi: 10.1186/1476-4598-9-1

27. Sivachandran N, Sarkari F, Frappier L. Epstein-Barr nuclear antigen 1 contributes to nasopharyngeal carcinoma through disruption of PML nuclear bodies. *PLoS Pathog.* 2008;4(10):e1000170.

doi: 10.1371/journal.ppat.1000170

28. Scaglioni PP, Yung TM, Cai LF, *et al*. A CK2-dependent mechanism for degradation of the PML tumor suppressor. *Cell.* 2006;126(2):269-283.

doi: 10.1016/j.cell.2006.05.041

29. Sivachandran N, Cao JY, Frappier L. Epstein-Barr virus nuclear antigen 1 Hijacks the host kinase CK2 to disrupt PML nuclear bodies. *J Virol.* 2010;84(21):11113-11123.

doi: 10.1128/JVI.01183-10

30. Holowaty MN, Zeghouf M, Wu H, *et al*. Protein profiling with Epstein-Barr nuclear antigen-1 reveals an interaction with the herpesvirus-associated ubiquitin-specific protease HAUSP/USP7. *J Biol Chem*. 2003;278(32):29987-29994.

doi: 10.1074/jbc.M303977200

31. Gruhne B, Sompallae R, Marescotti D, Kamranvar SA,

Gastaldello S, Masucci MG. The Epstein-Barr virus nuclear antigen-1 promotes genomic instability via induction of reactive oxygen species. *Proc Natl Acad Sci U S A*. 2009;106(7):2313-2318.

doi: 10.1073/pnas.0810619106

32. Cao JY, Mansouri S, Frappier L. Changes in the nasopharyngeal carcinoma nuclear proteome induced by the EBNA1 protein of Epstein-Barr virus reveal potential roles for EBNA1 in metastasis and oxidative stress responses. *J Virol*. 2012;86(1):382-394.

doi: 10.1128/JVI.05648-11

33. Gianti E, Messick TE, Lieberman PM, Zauhar RJ. Computational analysis of EBNA1 "druggability" suggests novel insights for Epstein-Barr virus inhibitor design. *J Comput Aided Mol Des*. 2016;30(4):285-303.

doi: 10.1007/s10822-016-9899-y

34. Bouckaert RR, Frank E, Hall M. *WEKA Manual for Version 3-9-1*. Hamilton, New Zealand: University of Waikato; 2016:1-341.

35. Holmes G, Donkin A, Witten IH. Weka: A Machine Learning Workbench. In: *Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference*. IEEE; 1994:357-361.

36. Kononenko I, Hong SJ. Attribute selection for modelling. *Future Gener Comput Syst*. 1997;13(2):181-195.

doi: 10.1016/S0167-739X(97)81974-7

37. Hall MA. *Correlation-based Feature Subset selection for Machine Learning*. Thesis. University of Waikato; 1988.

38. Hall M, Guetlein M. *BestFirst*; 2019. Available from: https://weka.sourceforge.io/doc.dev/weka/attributeselection/bestfirst.html [Last accessed on 2024 Nov 07].

39. Hall M. *GreedyStepwise*; 2019. Available from: https://weka.sourceforge.io/doc.dev/weka/attributeselection/greedystepwise.html [Last accessed on 2024 Nov 07].

40. Vujović Ž. Classification model evaluation metrics. *Int J Adv Comput Sci Appl*. 2021;12(6):599-606.

doi: 10.14569/IJACSA.2021.0120670

41. Ratner B. The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Target Meas Anal Mark*. 2009;17(2):139-142.

doi: 10.1057/jt.2009.5

42. Tatachar AV. Comparative assessment of regression models based on model evaluation metrics. *Int Res J Eng Tech*. 2021;8(9):853-860.

43. Gill J, Moullet M, Martinsson A, *et al*. Evaluating the performance of machine-learning regression models for pharmacokinetic drug-drug interactions. *CPT Pharmacometrics Syst Pharmacol*. 2023;12(1):122-134.

doi: 10.1002/psp4.12884

44. Damodharan S, Reddy SV, Sarojamma B. WEKA models for rainfall data. *Int J Emerg Technol Innovat Res*. 2022;9:C111-C119.

45. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22(1):69-77.

doi: 10.1002/qsar.200390007