Machine learning based QSAR model for therapeutically active candidate drugs with thiazolidinedione (TZD) scaffold

Irina GHOSH [1] (iD) **, Komal SINGH** [1] (iD) **, Venkatesan JAYAPRAKASH** [1] (iD) **, Sudeepan JAYAPALAN** [2]* (iD)

[1] Department of Pharmaceutical Sciences & Technology, Birla Institute of Technology, Mesra, 835215, Ranchi, India.
[2] Department of Chemical Engineering, Birla Institute of Technology, Mesra, 835215, Ranchi, India.

* Corresponding Author. E-mail: sudeepan@bitmesra.ac.in (S.J.); Tel. +91-9572635441.

ABSTRACT: Diabetes is a multifactorial metabolic disorder occurs due to uncontrolled persistent hyperglycaemia. The α-glucosidase enzyme plays an important role in management of diabetes. The α-glucosidase enzyme gets secreted by the brush border cells of small intestine which helps in converting maltose into glucose and thereby inhibiting the enzyme will help in lowering blood glucose level. In the present study, 100 compounds were selected having activity against α-glucosidase enzyme and they were used to build a machine learning based quantitative structure activity relationship model (QSAR). All the compounds were having thiazolidinedione (TZD) as the common nucleus. The molecules selected were divided into training and testing datasets of 80:20 ratio for various model development. The important molecular descriptors which will affect the target were chosen using recursive feature elimination (RFE) algorithm. The predictive models were created using machine learning regression techniques including Support Vector Regression (SVR), Random Forest Regression (RFR), Decision Tree Regression (DTR) and Gradient Boosting Regression (GBR). A comparison-based analysis was done between the various machine learning algorithms. The GBR and RFR gave the best R^2 value of 0.9992 and 0.9514 for the training dataset and 0.9414 and 0.8760 for the testing dataset respectively, followed by SVR and DTR. Thus, it concludes that the four-machine learning algorithm generates a highly predictive model for the unique compounds and a superior prediction capability for building a QSAR model for α-glucosidase enzyme inhibitors.

KEYWORDS: Machine Learning; α-glucosidase; TZD; Bioactivity; QSAR.

1. INTRODUCTION

Diabetes is one of the most common non-communicable diseases happening globally. Almost 80% of people suffering from diabetes live in developing countries like Indian subcontinent and China [1]. Diabetes happens when the body cannot efficiently use insulin, produce enough insulin, or both. Some of the processes involved in the onset of diabetes include the autoimmune destruction of the pancreatic β cells and the aberrant metabolism of protein, fat, and carbohydrates [2]. The number of people with diabetes worldwide has more than doubled during the last two decades. Diabetes mellitus is of different types, type 1, type 2 and gestational diabetes [3]. Type 2 diabetes (T2DM) affects more than 90% of diabetic people.

The emergence of type 2 diabetes in younger age groups, such as children, teenagers, and young people, is a worrying trend within the rapid growth. Enhancing the activity of two digestive system-located enzymes, intestinal α-glucosidase, and pancreatic α-amylase, is one method of managing diabetes [2]. The α-glucosidase enzyme plays a vital role in breaking down complex carbohydrates into simple sugars that can be absorbed and utilized for energy, which is essential for normal physiological functions. However, excessive activity of the enzyme can lead to decreased glucose absorption, causing significant problems for patients with type 2 diabetes [4]. To address the issue, α-glucosidase inhibitors have been developed and used to regulate glucose levels in type 2 diabetes mellitus. Several types of α-glucosidase inhibitors, including acarbose, miglitol, and voglibose, have been clinically applied for medicinal purposes to inhibit the activity of α-glucosidase [3]. However, each of these medicines is linked to certain major side effects, including increased food consumption, cardiovascular disease mortality, gastrointestinal pain, and weight gain, among others [5]. Currently, indigenous flora or their bioactive substances are employed to treat

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

hyperglycaemia via a variety of different mechanisms of action. These herbal treatments are frequently regarded as having no negative effects [6]. The protein PPARɣ, which is recognized for controlling the transcription of the insulin responsive gene, which is involved in the regulation of glucose production, transport, and subsequently utilisation, is modulated by thiazolidinediones [7]. There are currently several novel insulin sensitizers being researched.

The thiazolidinediones (TZD), also referred to as "glitazones," help type 2 diabetic patients better control their metabolic processes. Their impact on reducing blood sugar is mediated by an increase in insulin sensitivity. These substances are well-known for their capacity to lessen insulin resistance in adipose tissue, including muscle and the liver [8]. PPARɣ is a nuclear receptor that is activated by TZDs. The transcription of several genes involved in lipid and glucose metabolism is changed by the TZD-induced activation of PPARɣ [7]. This contains the genes that produce the adipocyte fatty acid binding protein, glucokinase, fatty acid transporter protein, fatty acyl-CoA synthase, lipoprotein lipase, and the GLUT4 glucose transporter [9]. Machine learning is a discipline of Artificial intelligence which focuses on the utilization of data and algorithms. During the history of rational drug development, a variety of machine intelligence techniques have been used to cut costs and reduce the time-consuming nature of traditional studies. Quantitative structure-activity relationship (QSAR) modelling is one of many machine-learning technologies that have been created over the past few decades that can swiftly and affordably discover possible biologically active molecules from thousands of candidate compounds [10]. To deal with the enormous volumes of data produced by contemporary drug development methods, deep learning approaches—which are more potent and effective—evolved into machine learning approaches when drug research entered the era of "big" data [11]. A common computational strategy called virtual screening (VS) is frequently used to direct rational drug discovery [12]. In the past, different QSAR models for VS have been produced using machine-learning techniques, which constitute one of the most crucial elements of artificial intelligence. All rational drug discovery techniques use the same QSAR modelling process. With the advancements in modelling techniques and the creation of descriptors, QSAR is frequently used throughout the preclinical study process [13]. All QSAR models created so far are built on the original QSAR premise that "similar substances have comparable actions." Nevertheless, even though various descriptor types and machine-learning techniques used for QSAR modelling each have their own advantages and disadvantages, the resulting models still experience the same problems, such as overfitting and active cliffs, which makes it impossible to predict new compounds, particularly those with chemical structures that differ from those in the training sets used to create QSAR models [14]. Wang et al., [15] studied the QSAR models on PPARɣ binding affinity using various machine learning algorithms. They found that the network like similarity graphs are positively correlated with the cross-validation of the models and also they found that the developed regression models could be used for the evaluation of PPARɣ based QSAR models. Further, Saxena et al., [16] studied the insulin resistance metagenes of type 2 diabetes using variousmachine learning methodology such as support vecotr machine, XGBoost, Random Forest and so on. They found that the developed model gave the result of 73% accuracy across 64 human adipose tissue samples. Similarly, various researchers developed different machine learning based QSAR models for different targets [17-22]. As a result, fresh initiatives are being made to include new modelling tools into QSAR to make it more suitable for drug discovery.

In this study a Machine Learning based QSAR model was developed for compounds having thiazolidinedione scaffold targeting inhibition of α-glucosidase enzyme using four different machine learning algorithms namely Support Vector Regression (SVR) [23], Gradient boosting Regression (GBR) [24], Random Forest Regression (RFR) [25] and Decision tree Regression (DTR) [26]. The model's efficiency was evaluated by using coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

2. RESULTS AND DISCUSSION

A ML-based predictive model of anti-diabetic compounds acting on inhibition of α-glucosidase enzyme having TZD scaffold were studied. The predictive models were built using various machine learning algorithm.

The machine learning predictive model for α-glucosidase inhibitors having TZD scaffold was developed by utilizing the top 50 features by recursive feature selection. Accordingly, four machine learning based regression algorithms were used such as Support Vector Regression (SVR), Decision Tree Regression (DTR), Gradient boosting Regression (GBR), and Random Forest Regression (RFR).

The model's performance was evaluated by using various statistical evaluation metrics like coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Hyperparameters were used in the building of the model which helped in controlling the behaviour of the model during

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

training and testing prediction. The hyperparameters used in the RFR is 'max_depth' , 'max_features' and 'n_estimators', for SVR it is 'c', 'gamma', and 'kernel', for GBR it is 'max_depth', 'min_samples_leaf' and 'n_estimators', for DTR it is 'max_depth', 'min_samples_leaf' and 'min_samples_split'. The likelihood of estimating actual data from the regression line is provided by R^2 and its value lies between 0 and 1. The effectiveness of estimate increases as the R^2 value tends approaching 1. The magnitude of the mistake in the value predictions is measured by MAE and RMSE estimates. The better the MAE and RMSE numbers, the more accurate is the prediction model.

2.1. Model Accuracy Analysis for α-Glucosidase Inhibitors

The performance of the model was evaluated based on the metrics R^2, RMSE and MAE contingent to SVR, RFR, DTR and GBR algorithmic model. The R^2, RMSE and MAE values proves that the model is performing accurately and precisely. Out of all the models, the training dataset in the GBR model is the most suitable one with having R^2 value 0.9992, MAE value 0.0078, and RMSE value 0.0159, thus a suitable choice as a α-glucosidase inhibitor; followed by RFR, SVR and DTR model. In the case of test dataset also, GBR model gave better performance with R^2 value 0.9410, MAE value 0.2692, and RMSE value 0.4236; followed by RFR, SVR and DTR. The values are given in Table 1.

Table 1. Performance Evaluation Metrics of different models after fine-tuned hyperparameters for the training and testing set.

	SVR		RFR		DTR		GBR	
Evaluation Metrics	Train	Test	Train	Test	Train	Test	Train	Test
R^2	0.9445	0.8576	0.9514	0.8760	0.5697	0.4670	0.9992	0.9410
MAE	0.3196	0.3530	0.1443	0.3976	0.2488	0.3738	0.0078	0.2692
RMSE	0.5502	0.4536	0.2181	0.5509	0.4022	0.5785	0.0159	0.4236

2.1.1. Support Vector Regression (SVR) analysis

In Figure 1 (a) and Figure 2 (a), the scatter plot was studied for the detailed examination of the SVR model's accuracy and performance in both training and testing dataset. The accuracy of the model is defined by trajectory of data near the slope line at 45^0. In Figure 1 (a). the training data were marginally diverted from the slope line. This behaviour of the values was also casted in the R^2 precision indices. While in the Figure 2 (a), the test data is predominantly more diverted from the slope line.

An additional graphical examination was done enlisting the Observed and Predicted values for different Experimental runs of SVR model, after fine-tuned hyperparameters for the training and test dataset for the prediction of pIC50 of α-glucosidase enzyme inhibitor molecules is shown in Figure 3 (a) and Figure 4 (a). In Figure 3 (a), the observed and predicted values for the training dataset is almost concurrent. And for the test dataset in Figure 4 (a), the observed and predicted values are slightly diverged from each other; their performance parameter is also inscribed in Table 1.

A residual plot was sketched for SVR models after fine-tuned hyperparameters for the training and test dataset for the prediction of pIC50 of α-glucosidase enzyme inhibitor molecules is shown in Figure. 5 (a). In the Figure 5 (a) graph the test and training values are close to the zero line and a symmetry is followed between the values is found; thus, it shows that the SVR model is a good fit for the data points.

2.1.2. Random Forest Regression (RFR) analysis

In Figure 1 (b) and Figure 2 (b), a Scatter plot analysis for RFR model was done after fine-tuned with hyperparameters for the training and testing dataset for the prediction of pIC50 of α-glucosidase enzyme inhibitor molecule respectively. For the training dataset (Figure 1 (b)), the values are much more aligned to the slope line in comparison to the testing dataset. In the testing dataset (Figure 2 (b)), the data are clustered near the slope line, thus no proper correlation between the data point was found.

In Figure 3 (b) and Figure 4 (b), the observed and Predicted values for different Experimental runs of RFR model after fine-tuned hyperparameters for the training and test dataset for the prediction of pIC50 of α-glucosidase enzyme inhibitor molecules. In Figure 3 (b), for the training dataset the observed and predicted values are harmonized with R^2 value 0.9514. While for the test dataset in Figure 4(b), the values are slightly scattered along the observed and predicted line with R^2 value 0.8760.

In Figure 5 (b), a residual plot analysis for RFR model was done. The values are equitably distributed across the zero line, which indicates that the RFR model is a perfect fit for both the training and test dataset.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

2.1.3. Decision Tree Regression (DTR) analysis

A scatter plot analysis of DTR model was done in Figure 1 (c) and Figure 2 (c) for the training and test dataset. The data points in training and test dataset are not following a synchronous trend along the 450 slope line. Thus, providing an inferior fit compared to other models.

In Figure 3 (c) and Figure 4 (c), the observed and predicted values for different experimental runs of DTR model after fine-tuned hyperparameters for the training and test dataset for the prediction of pIC50 of α-glucosidase inhibitor molecules are shown. In Figure 3 (c), for the training dataset the predicted and observed line is marginally correlated while for the test set in Figure 4 (c), the predicted and observed line is irregularly correlated. Their respective R² values are 0.5697 and 0.4670. Thus exhibits an inferior model accuracy.

In Figure 5 (c), a residual plot analysis was done, the predicted values are equally distributed across the zero line and a symmetry was maintained. Thus, indicating a good coherence in the model.

2.1.4. Gradient Boosting Regressor (GBR) analysis

Figure 1 (d) and Figure 2 (d) shows scatter plot of GBR model after fine-tuned hyperparameters for the training and testing dataset for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules. In Figure 1 (d), the values are packed along the slope line and thus the GBR model gives the best fit with all the actual variables. In Figure 2 (d), it has been observed that in the test dataset, the values are proportionally distributed across the zero line.

In Figure 3 (d), the observed and predicted plot for training dataset, where the observed values are totally synchronous with the predicted value having R² value 0.9992. Whereas in Figure 4 (d), the observed and predicted values are marginally diverse with R² value of 0.9410.
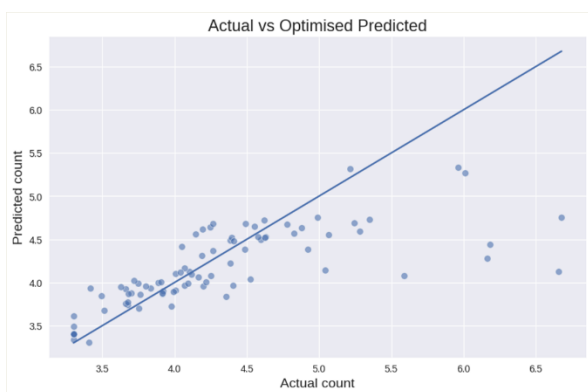
Figure 5 (d) inscribes the residual plot for GBR model. The values are almost near to zero line thus indicating a better fit compared to other models.
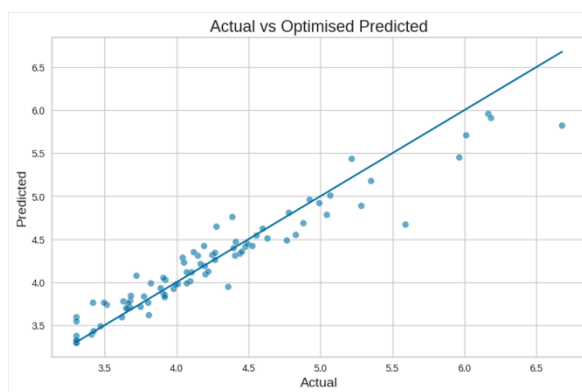
2.1.5. Validation of Best Generated Models

The four different ML models of SVR, RFR, DTR and GBR were developed using Python codes. The dataset were splited into 80:20 ratio and the different models were trained as well as tested. GBR model shows the better performance compared with other different four models. Somes of the already reported compounds were validated using developed GBR model [28, 29]. The performance of the validated results for the selected compounds are shown in Table 2. It is seen from the Table 2, the randomly selected two compounds shows an error of 1.08% and 4.23%.

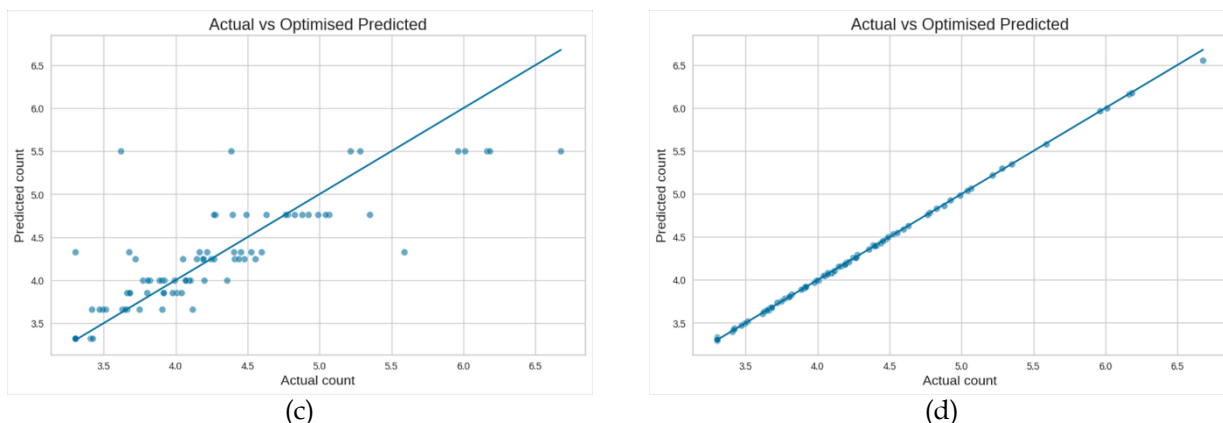**Table 2**. Performance Evaluation of validated results of best performed GBR model.

Smiles	pIC50	Standardized pIC50	Predicted Standardized pIC50	Error (%)
CCCCCCSc2nc1ccccc1n2CC(=O)NN3C(=O)CSC3c4cccc(OC)c4O	4.6179	0.3604	0.3565	1.08
O=c2[nH]c(=Nc1ccc(N(=O)=O)cc1)sc2=Cc4cn(c3ccccc3)nc4c5ccccc5	4.0347	0.3822	0.3984	4.23

(a)

(b)

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold
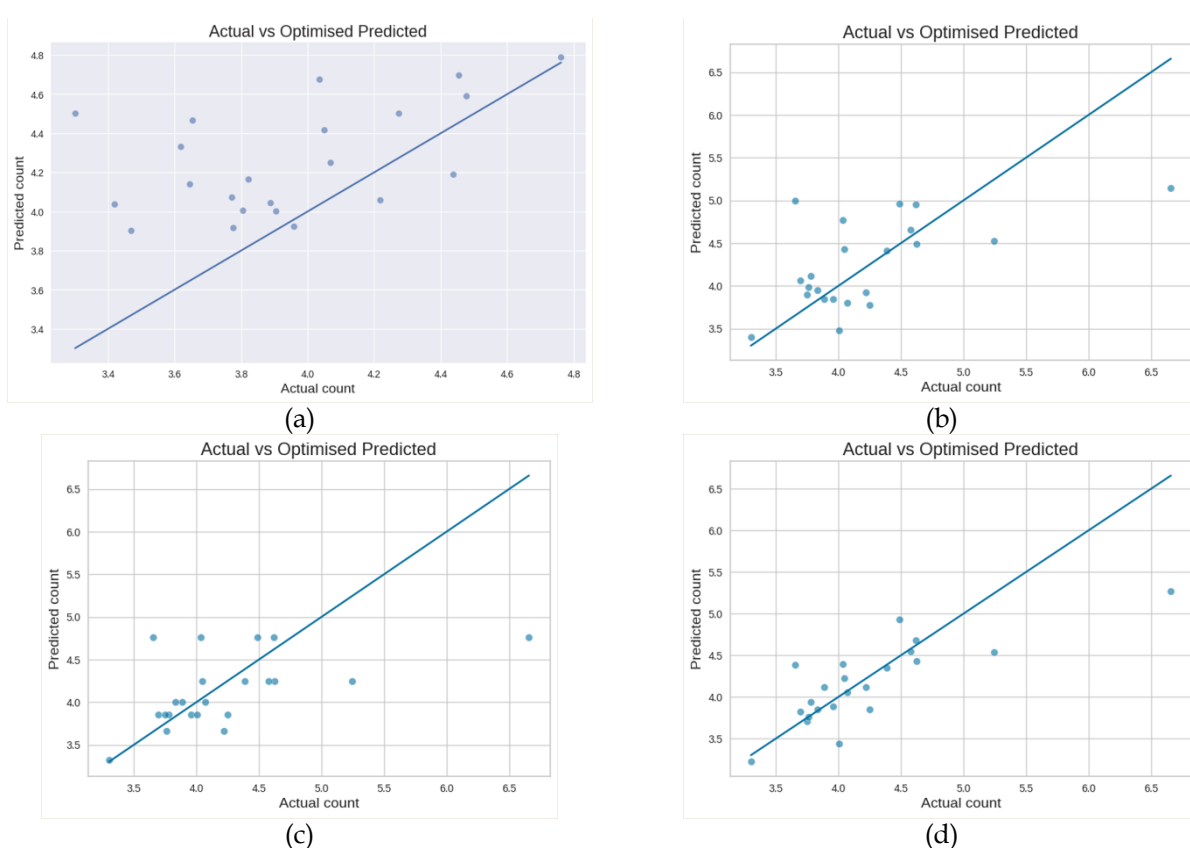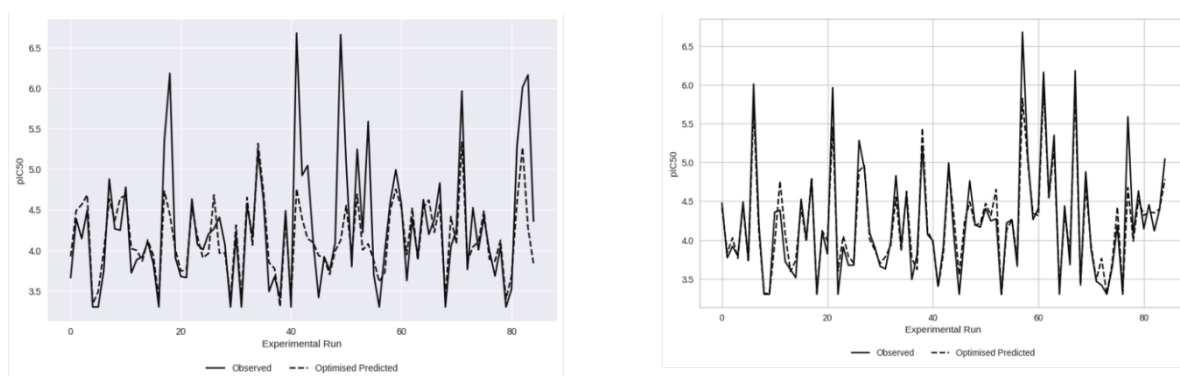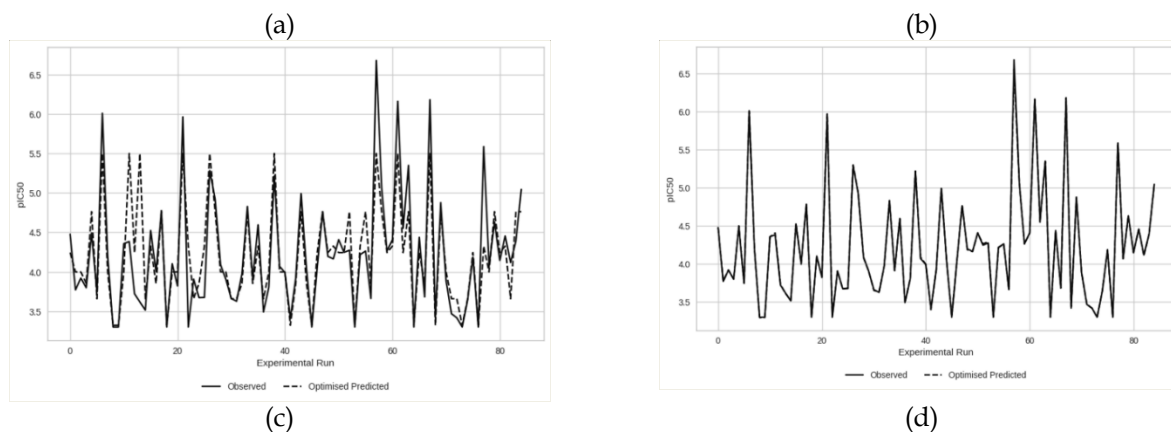
Journal of Research in Pharmacy

Research Article

Figure 1. Scatter plot for the training dataset for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules (a) SVR (b) RFR (c) DTR and (d) GBR
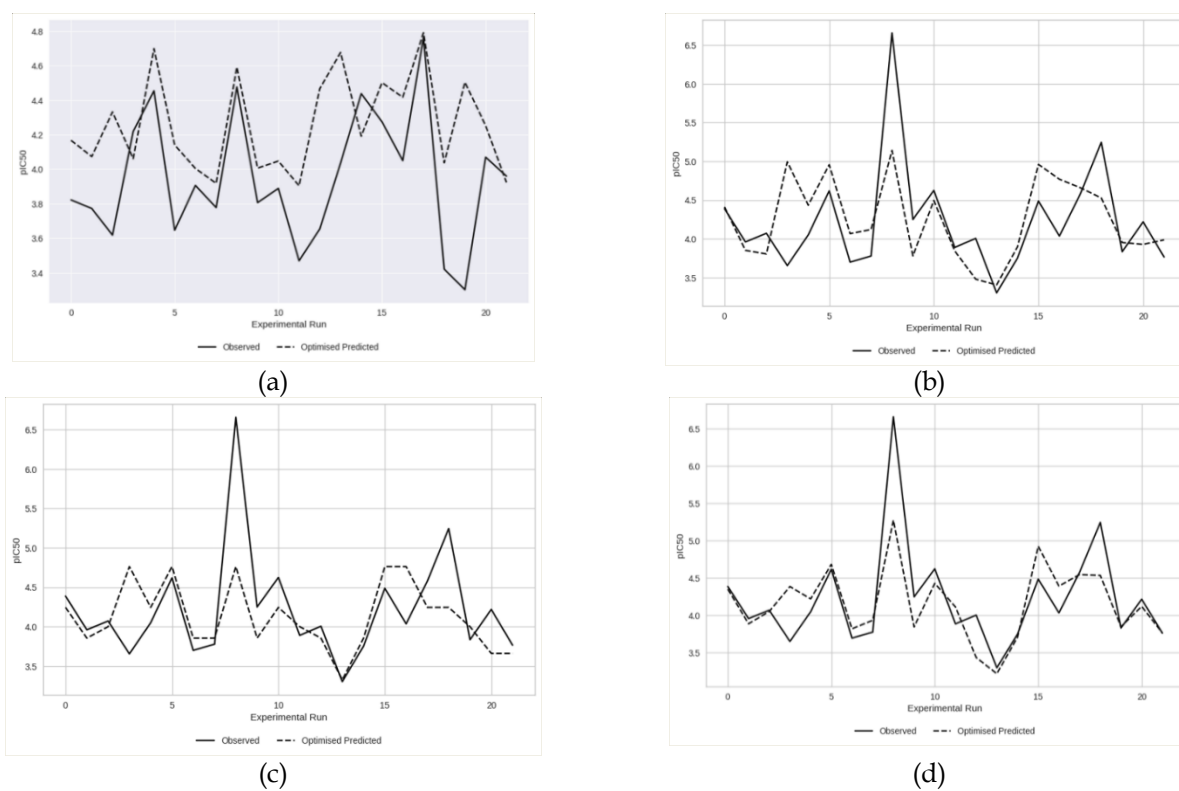

Figure 2. Scatter plot after for the test dataset for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules (a) SVR (b) RFR (c) DTR and (d) GBR

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

(a)

(b)

(c)

(d)

Figure 3. Observed and Predicted values for different Experimental runs for the training dataset for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules (a) SVR (b) RFR (c) DTR and (d) GBR

(a)

(b)

(c)

(d)
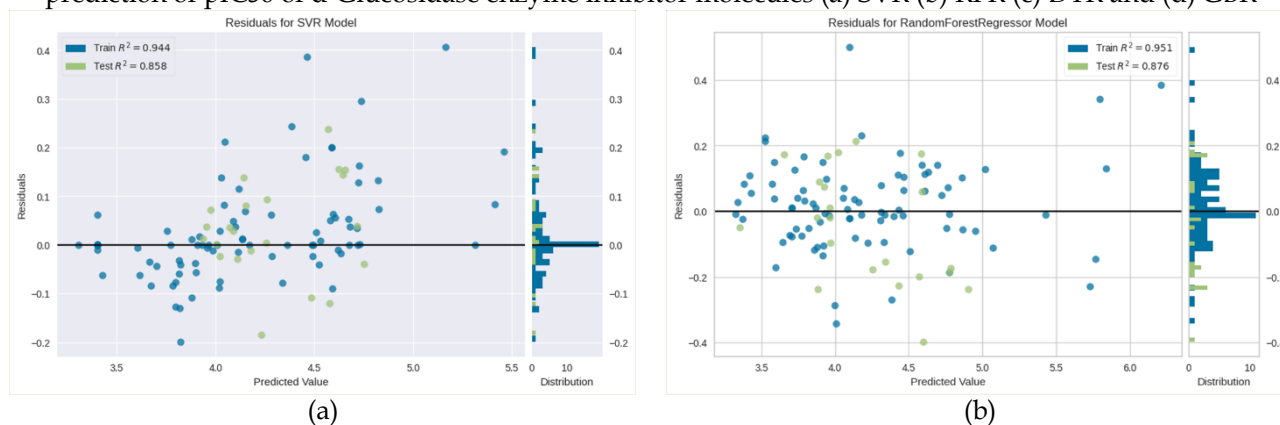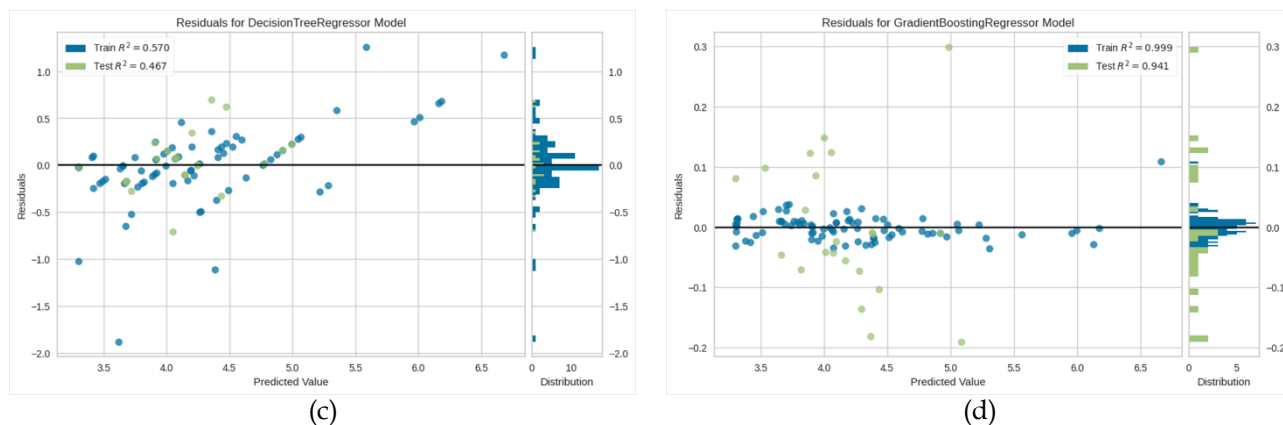
Figure 4. Observed and Predicted values for different Experimental runs for the testing dataset for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules (a) SVR (b) RFR (c) DTR and (d) GBR

(a)

(b)

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

Figure 5. Residual plots for the prediction of pIC50 of α-Glucosidase enzyme inhibitor molecules (a) SVR (b) RFR (c) DTR and (d) GBR

3. CONCLUSION

Classical machine learning approaches has developed into a variety of alternative strategies, including OECD-QSAR, Pep-QSAR, hybrid QSAR and it is still a popular tool for studying areas like drug design or drug related ventures. Nowadays many drugs are getting discovered with the help of computational methods, which eventually lead to cut down in costs and speed up the process of discovery. It also reduces the lots of physical work which used to be needed before. In this project a ML-based QSAR model was developed for thiazolidinedione scaffold. The model was built using four different types of machine algorithm likely SVR, RFR, DTR and GBR. The dataset was built by collecting compounds from various scientific articles, consecutively the descriptors were generated by using PaDEL software, and then the important features were selected by Recursive feature Elimination tool. The four models were built and then in between them a correlative assessment was performed with the help of statistical measure likely R^2, MAE and RMSE. The R^2 value obtained by GBR and RFR is 0.9410 and 0.8760 for test dataset, while for training set it is 0.9992 and 0.9514 respectively, thus proving to be the most suitable algorithm for model building. Followed by SVR and DTR with R^2 value 0.8576 and 0.4670 for test dataset, while for training set it is 0.9445 and 0.9514 respectively. From the validation of randomly selected two compounds shows the error of 1.08% and 4.23% from the developed and best performed GBR model among other developed models. Thus, we can clearly draw the result, that is GBR and RFR can be used for building a QSAR model for anti-diabetic agents acting on inhibiting α-glucosidase enzyme.

4. MATERIALS AND METHODS

4.1. Data collection

- The QSAR model was developed using google Colab notebook [27].
- In this study, 100 molecules having thiazolidinediones (TZD) as the central nucleus acting on a same target which is α-glucosidase enzyme are collected from various scientific articles for the above target [2, 3, 6, 28-33].
- The molecules used for building the dataset should have their disease set common, must be acting on similar target enzyme/receptor, identical scaffold, the activity details, and similar assay procedure.
- In the dataset the Scaffold is- Thiazolidinedione, Target is - α-glucosidase enzyme, Activity taken- IC50 and Assay procedure used is - α-glucosidase inhibition assay.
- Inhibitors having IC50 values were extracted from the data.
- SMILES were generated for the collected compounds using Molinspiration [34].
- Using various machine learning approaches, the entries were used to construct predictive models for each target.
- The equation pIC50 = -log10(IC50), where IC50 was in molar concentration (µM), the equation was used to convert the half-maximal inhibitory concentration (IC50) of these unique entries to pIC50.The dataset used for model development is provided in Supplementary Table S1 for α-glucosidase enzyme inhibition.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

4.2. SMILES generation

The SMILES were generated for encoding the molecular structure of the compounds. The activity data for the compounds which is IC50 converted into pIC50. The Molinspiration software was used for the SMILES generation [34]. Then, they were given a unique id for every compound on the dataset.

Molinspiration is a Java based cheminformatics software. It has many tools and packages in it which helps in generation of various molecular properties needed in QSAR modelling. Molinspiration have mib engine which helps in molecular processing, SMILES depiction etc. In Molinspiration, the molecular structure of the compound which acts as the input are drawn initially. Then, the SMILES (output) give us connectivity of the atoms present in the compound. The vice-versa can also be done using Molinspiration software. The dataset along with their SMILES is given in Supplementary Table S1.

4.3. Molecular Descriptors generation

Molecular descriptors are representation of quantitative properties of a molecule. They are used to characterize the chemical structure of the compounds and predict their various chemical and biological processes. Molecular descriptors can be calculated using various computational tools. These tools will analyse the molecular structure, including its size, shape, electronic properties, and chemical composition and then will generate the descriptors. There are various types of chemical descriptors like 1D, 2D, and 3D [35].

In this study, 2D descriptors are considered. 2D descriptors are mathematical representation of 2-dimensional molecular structure of a compound. The various types of 2D descriptors are topological, geometrical, and electronic. The topological descriptors give information about connectivity in between the atoms in the molecule, geometrical descriptors represent molecular size and shape. Electronic descriptors give information about electronic properties likewise charges etc. The examples of 2D descriptors are number of atoms, the number of bonds, the molecular weight, the number of rings, the number of hydrogen bond donors and acceptors and the LogP. 2D descriptors are used in drug design, building QSAR model and virtual screening [36].

For building QSAR based prediction model for α-glucosidase enzyme (target), an open source PaDEL-descriptor software [37] was used to generate the molecular descriptors. The chemical descriptors are calculated for every molecule present in the dataset [38]. The information regarding molecule structure, such as molecular weight, the number of bonds, the solvent accessible area, etc., is depicted by these molecular descriptors including fingerprints. According to their dimensionality, the descriptors are divided into 1D, 2D, and 3D features and are essential for comprehending the quantitative structure-activity relationship of molecules. All the descriptors generated are reported in Supplementary Table S2.

4.4. Feature selection

The process of picking relevant characteristics or variables from a wider set of features that are present in a dataset is known as feature selection. The feature selection process is done to improve the performance of ML-model by minimising the number of pointless or redundant features utilised during training of the model [39]. It helps in reducing overfitting, mitigating the curse of dimensionality. Feature selection helps in improving the model's interpretability and helps in selection of the most important features from a dataset. Feature selection helps in cutting down the computational cost.

Recursive feature elimination (RFE) is one of the feature selection techniques used in machine learning to rank the features, aims in selecting the most significant feature in a dataset by repeatedly removing the least significant features from a dataset until the desired number of features is obtained.

The primary objective of RFE feature selection technique is to reduce the complexity of the model which in turn enhance the model performance. Also, RFE method prevents overfitting by removing unnecessary noise generated from less significant features for the model. Briefly, using RFE feature selection technique enhances the model performance and interpretability through the elimination of one feature or small set of features at a time [40].

The RFE method starts with the:
- Initial training of a regression model using all considered features in the dataset.
- After the initial training of the model, the significance of each feature is ranked and assessed using coefficients or feature_importance_ attribute for regression models.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

- RFE recursively eliminates a least significant features per iteration based on removing any existing dependencies and collinearities in the model. The iteration process will be continued until specified number of features are reached.
- Cross-validation will split the dataset into number of subsets. One set of data is used for testing and the remaining data is used for training the model. Again, each time a different set of data will be used for testing and the remaining for training the model. The iteration process is repeated until the specified task is reached. In this way, cross-validation will help RFE to repeat the iteration process with different set of data's each time to evaluate the algorithm efficiently.
- The features which are most significant are extracted from the dataset to predict the target variable efficiently.
- In general, RFE uses two parameters such as number of features to be selected and the choice of model can also be specified by the user.
- Finally, the model will be re-evaluated using the most significant features in the dataset, resulting in an enhancement of model efficiency [40].

Steps followed in the present work for implementing Recursive feature elimination-

- Firstly, the necessary packages are imported from sklearn.feature_selection import RFE in the algorithm.
- Then load the input and output variables of various features for the study.
- Initiate RFE process by training a regression model with all considered features in the dataset and mention the number of features to be selected and the number of steps for cross-validation. This work utilised top 50 features and 5 steps are selected for the study.
- After the initial training, the features are ranked and the feature importances are assessed using feature_importance_ attribute for tree based regression model.
- The pertinent top 50 features were chosen from the 10,851 features to serve as input variables for model training purpose using RFE. To prevent overfitting and the constraint of dimensionality, feature selection is essential. The selected top 50 features of each α-glucosidase inhibitor for each technique have been given Supplementary Table S3. The top 50 ranked features were used for further study.

4.5. Machine learning methods

The model was build using predictive algorithm for α-glucosidase enzyme target using four different machine learning methodology viz., Support Vector Regression (SVR), Random Forest Regression (RFR), Decision Tree Regression (DTR) and Gradient Boosting Regression (GBR).

4.5.1. Support Vector Regression (SVR)

A supervised machine learning method for tackling regression issues is support vector regression (SVR). When examining the link between a dependant variable and one or more predictor variables, regression analysis is helpful. To learn a regression function that maps from input predictor variables to output observed response values, SVR formulates an optimisation problem [41]. SVR has a good performance for processing high-dimensional data and strikes a compromise between model complexity and prediction error [42].

4.5.2. Random Forest Regression (RFR)

A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting. It helps in regression tasks which includes prediction of a continuous numerical variable. For using random forest regressor, a set of input features must be provided and their associated target values. RFR can handle many features [43].

4.5.3. Decision Tree Regression (DTR)

A decision tree creates tree-like models for classification or regression problem. It incrementally develops an associated decision tree while segmenting a dataset into smaller and smaller sections. The outcome is a tree containing leaf nodes and decision nodes. Two or more branches, one for each value of the tested characteristic, can be found on a decision node. A choice regarding the numerical aim is represented

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

by a leaf node [44]. The root node is the topmost decision node in a tree and corresponds to the best predictor. Both category and numerical data can be processed using decision trees [45].

4.5.4. Gradient boosting Regression (GBR)

It is a well-liked machine learning approach, utilised for both classification and regression problems. It is an ensemble method that creates a powerful prediction model by combining several weak models [46].

Gradient boosting's core principle is to iteratively add fresh weak models to the ensemble and then modify the weights of the samples in accordance with the mistakes. Each new model is specifically trained to forecast the residuals, or the discrepancies between true and anticipated values, of the older models.

4.6. QSAR model building

A QSAR model predicts the activity or property of a compound based on its structural and chemical properties. Methods required for model building- [47].

- A dataset must be prepared containing set of compounds with known activities and characteristics. The database should contain a variety of chemical structures that cover the relevant chemical space.
- The data should be cleaned, pre-processed and be transformed into csv form which could be easily understood by the data analysis and machine learning tool.
- The descriptors were generated and out of them important features were to be selected.
- The selected features were split into training set and test set.
- After that model was developed using various algorithm like SVR, DTR, RFR and GB [48].
- The hyperparameters were used for the tuning of the model for better performance.
- The model was evaluated on the test data by using various statistical measures like coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

4.7. Hyperparameters

Hyperparameters are parameters that are set by the user before training a machine learning model rather than being learned from the data during training. These variables regulate how the learning algorithm behaves and can greatly affect how well the model performs [49].

Hyperparameters example includes learning rate, hidden layers number, batch size, neurons present in each layer, activation functions and regularization of strength, etc.

4.7.1. Hyperparameter Tuning

Hyperparameter tuning is the method of choosing the ideal values for a machine learning model's hyperparameters, conducive to get the best performance on a particular task or dataset. Grid search is a popular method for hyperparameter tuning, which entails defining a range of values for each hyperparameter and then assessing the model's performance on a validation set for each set of hyperparameter values in the search space [50].

The learning rate, the maximum depth, the number of estimators, and the minimum number of samples necessary to split an internal node are the four hyperparameters that this dictionary provides for a tree-based machine learning model. A selection of potential values is given for each hyperparameter.

These hyperparameters could be used in a grid search strategy to train a model, which compares all feasible combinations of hyperparameters to determine which combination performs the best on a validation set [51].

4.7.2. Hyperparameters in SVR

The hyperparameters used in the RFR is 'C,' 'Gamma' and 'Kernel' -
- 'C' - Support vector machines (SVMs) are frequently used in machine learning to control the trade-off between attaining a low training error and a low testing error. [52]
- The penalty for incorrectly classifying training instances is set by the regularization parameter C. More misclassifications in the training set will be permitted by a lower value of C, which could lead to a more straightforward model with higher bias and lower variance [53]. A bigger value of C, on the other hand, will result in a more complex model with reduced bias

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

and higher variance since it will impose a stronger penalty for misclassifications. The values selected for hyperparameter tuning is [1,10,100], while the fine-tuned value came is 1.

- 'Gamma' – The hyperparameter gamma is defined by impact of a single training example. It governs the flexibility of the decision border by deciding the bandwidth of the kernel function.[54]
- The trade-off between overfitting and underfitting the data is controlled by the gamma value. A decision boundary with a more complex form will be produced by a larger gamma value, which may cause overfitting [55]. However, a lower gamma value will produce a smoother decision boundary, which may result in underfitting. The selected value for hyperparameter gamma is [1,0.1,0.001], the fine-tuned value came is 0.01.
- 'Kernel'- The type of kernel function to be utilized in a machine learning method is determined by a kernel hyperparameter. Several kernel functions, including linear, polynomial, and radial basis function (RBF) kernels, can be selected using the kernel hyperparameter [56].
- In machine learning, picking the appropriate kernel hyperparameter is crucial because it has a significant impact on the algorithm's performance[57]. The algorithm's accuracy and efficiency may be impacted by the kernel function selection, which may need to be adjusted to produce the best results for a given dataset.

4.7.3. Hyperparameters in RFR

The hyperparameters used in the RFR is 'max_depth' , 'max_features' and 'n_estimators'.

- 'max_depth' – 'max_depth' – The max_depth hyperparameter determines the level of decision nodes allowed in each tree. Hyperparameter tuning like grid search is used to find the optimal value for max_depth [58]. The optimal value will vary depending on the dataset and the issues arises.
- 'max_features' – 'max_features' – This helps in analysis of text like document classification or sentiment analysis. It helps in determining the maximum features that are extricated from text data and to be used as input for the model [59]. 'max_features' help in reducing the dimensionality of data and avoid overfitting. Therefore, it is crucial to select a suitable value for max features based on the size of the dataset and the difficulty of the task. Important data may be lost if max features is set too low, and if it is set too high, overfitting and a slower model training process may result [60].
- 'n_estimators' – The n_estimators hyperparameter determines the size of the ensemble. The RFR model may perform better when the value of n_estimators is increased, but at the expense of more computing complexity and longer training times [61].

4.7.4. Hyperparameters in DTR

The hyperparameters used in the DTR is 'max_depth', 'min_samples_leaf' and 'min_samples_split'.

- 'max_depth' – The max_depth hyperparameter controls the tree's maximum depth. Without a limitation on the depth, the tree can overcomplicate and overfit the training set, which would result in poor generalization to the test set [62]. Overfitting can be avoided, and the complexity of the tree can be managed by using the max_depth hyperparameter.
- 'min_samples_leaf' – The min_samples_leaf sets the least number of samples required to be at leaf node of the tree. Specifying value for min_samples_leaf helps to prevent overfitting and controls the depth of the tree and reduces the variance of developed model [63].
- 'min_samples_split' – It enumerates the minimum number of samples required to split an internal node. If samples present in a node is lesser than min_samples_split then the node does not split further and then it will become a leaf node. By guaranteeing that each internal node has sufficient samples to produce a trustworthy split, this can help prevent overfitting [64].
- Increasing the value for min_samples_split helps in building a simpler and easier understandable model; but accuracy gets suffered. Setting a smaller value, on the other hand, may result in overfitting as well as a more complex model with higher accuracy. The dataset and the issue that needs to be solved determine the ideal number for min samples split. Techniques like grid search can be used to fine-tune [65].

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

4.7.5. Hyperparameters in GBR

The hyperparameters used in the GBR is 'max_depth', 'min_samples_leaf' and 'n_estimators' –
- 'max_leaf_nodes' – The max_leaf_node hyperparameter regulates the maximum number of leaf node in each ensemble. It limits the complexity and prevents overfitting. Lowering the value of max_leaf_node aids in preventing overfitting [66].
- 'min_samples_leaf' - The optimal value for min_samples_leaf depends on the complexity of the problem, the amount of training data available, and the desired balance between model performance and interpretability overfitting. A greater score, nevertheless, can also indicate underfitting if the model is overly straightforward and fails to account for the complexity of the data [67].
- 'n_estimators' - Each tree in a gradient boosting ensemble is trained using the mistakes of the previous tree. Performance can also be enhanced by increasing n estimators, although if the value is too high, overfitting may result [68] . The best value for n_estimators is determined by the dataset and the problem handled [69].

4.7.6. Hyperparameter fine-tuned values

- The selected values for hyperparameter 'C', 'Gamma' and 'kernel' used in SVR model is [1,10,100], [1,0.1,0.01] and [rbf] and the fine-tuned values came are 1, 0.01 and rbf respectively.
- The selected values for hyperparameter 'n_estimators', 'max_features' and 'max_depth' used in RFR model is [10,100,500], [sqrt, log2] and [5,10,20] and the fine-tuned values came are 10, log2 and 10 respectively.
- The selected values for hyperparameter 'max_depth', 'min_samples_split' and 'min_samples_leaf' used in DTR model is [7,8,9], [7,8,9] and [7,8,9] and the fine-tuned values came are 7, 7 and 9 respectively.
- The selected values for hyperparameter 'max_leaf_nodes', 'max_depth' and 'min_samples_leaf' used in GBR model is [7,8,9], [6,7,8] and [3,4,5] and the fine-tuned values came are 9, 7 and 4 respectively.
- Table 3 mentioned the values of selected hyperparameter and fine-tuned hyperparameter.

4.8. Model evaluation

The performance of the models was assessed by calculating coefficient of determination (R^2), root mean absolute error (RMSE) and mean absolute error (MAE) The equation as follows-

4.8.1. Coefficient of determination (R^2)

It measures how much effectively a statistical model forecasts an outcome. The dependant variable in the model is a representation of the result. R^2 can have a value of 0 or 1, with 1 being the maximum achievable. If a model's R^2 is closer to 1, therefore it means that its predictions are accurate. R^2 is a more precise estimation of goodness of fit [70]. R^2 gives a proportion amount of variation in the dependant variable that the model can explain. The equation of R^2 is explained in Equation 1,2 and 3 [71], [26].

y_i - Actual value for the i-th data point

$\hat{y}$ – Predicted value.

$\bar{y}$- Mean value

$$RSS = \Sigma(y_i - \hat{y}_i)^2 \tag{1}$$

$$TSS = \Sigma(y_i - \bar{y})^2 \tag{2}$$

$$R^2 = 1 - \frac{RSS}{TSS} \tag{3}$$

RSS = Sum of square of residuals   TSS = Total sum of squares

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy
Research Article

Table 3. Fine-tuned hyperparameters for the prediction of pIC50 of α-glucosidase inhibitors molecules

Model	Hyperparameter	Selected Values	Fine-tuned Values
SVR	C	1, 10, 100	1
	Gamma	1, 0.1, 0.01	0.01
	Kernel	rbf	rbf
RFR	n_estimators	10, 100, 500	10
	max_features	sqrt, log2	log2
	max_depth	5,10,20	10
DTR	max_depth	7, 8, 9	7
	min_samples_split	7, 8, 9	7
	min_samples_leaf	7, 8, 9	9
GBR	max_leaf_nodes	7,8,9	9
	max_depth	6,7,8	7
	min_samples_leaf	3,4,5	4

4.8.2. Mean Absolute Error (MAE)

The effectiveness of a regression model is measured using the mean absolute error (MAE). It is described as the typical absolute difference between the model's projected values and the actual values of the underlying data [57]. When mistakes are distributed evenly across the data, the MAE, which reflects the mean magnitude of the model's errors in its predictions, is a helpful tool for assessing a model's performance. Since it is not sensitive to the existence of outliers, it is particularly helpful when the mistakes are symmetrically distributed and there are no extreme outliers. Here, n= number of observations, Eact = true value of the ith observation and Epred = predicated value of the ith observation. The bars represent the absolute value. Σ= Represents the sum of differences (As shown in Equation 4) [72], [26].

$$MAE = \frac{1}{n}\sum_{n=1}^{n}\left|E_i^{pred} - E_i^{act}\right| \tag{4}$$

4.8.3. Root Mean Absolute Error (RMSE)

One of the methods most frequently used to assess the accuracy of forecasts is root mean square error, also known as root mean square deviation. It illustrates the Euclidean distance between measured true values and forecasts. For evaluating a model's performance in machine learning, whether during training, cross-validation, or monitoring after deployment, it is very helpful to have a single number [59]. One of the most popular metrics for this is root mean square error. It is an appropriate scoring method that is simple to comprehend and consistent with some of the most widely used statistical presumptions. RMSE is the average of the squared difference between the model's predicted value and the actual value. Here, n= number of observations, Eact = true value of the ith observation and Epred = predicated value of the ith observation. The bars represent the absolute value. Σ= Represents the sum of differences (As shown in Equation 5) [73], [26].

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{n}\left(E_i^{pred} - E_i^{act}\right)^2} \tag{5}$$

Ghosh et al.
Machine Learning based QSAR model for therapeutically active candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

REFERENCES

[1] Zimmet PZ, Magliano DJ, Herman WH, Shaw JE. Diabetes: a 21st century challenge. Lancet Diabetes Endocrinol. 2014;2(1): 56-64. https://doi.org/10.1016/s2213-8587(13)70112-8.

[2] Fettach S, Thari FZ, Hafidi Z, Tachallait H, Karrouchi K, El Achouri M, Cherrah Y, Sefrioui H, Bougrin K, Abbes Faouzi ME. Synthesis, α-glucosidase and α-amylase inhibitory activities, acute toxicity and molecular docking studies of thiazolidine-2,4-diones derivatives. J Biomol Struct Dyn. 2022;40(18): 8340-8351. https://doi.org/10.1080/07391102.2021.1911854.

[3] Gaba S, Singh G, Monga V. Design, synthesis, and characterization of new thiazolidinedione derivatives as potent _ampersandsignalpha;-glucosidase inhibitors. Pharmaspire. 2021 Oct;13: 182-193. https://www.isfcppharmaspire.com/article_html.php?did=13767&issueno=0.

[4] Forouhi NG, Wareham NJ. Epidemiology of diabetes. Medicine (Baltimore). 2010;38(11): 602-606. https://doi.org/10.1016/j.mpmed.2010.08.007.

[5] Reichard P, Pihl M. Mortality and treatment side-effects during long-term ıntensified conventional ınsulin treatment in the Stockholm diabetes ıntervention study. Diabetes. 1994;43(2): 313-317. https://doi.org/10.2337/diab.43.2.313.

[6] Hussain F, Khan Z, Jan MS, Ahmad S, Ahmad A, Rashid U, Ullah F, Ayaz M, Sadiq A. Synthesis, in-vitro α-glucosidase inhibition, antioxidant, in-vivo antidiabetic and molecular docking studies of pyrrolidine-2,5-dione and thiazolidine-2,4-dione derivatives. Bioorg Chem. 2019;91:103128. https://doi.org/10.1016/j.bioorg.2019.103128.

[7] Zinman B, Gerich J, Buse JB, Lewin A, Schwartz S, Raskin P, Hale PM, Zdravkovic M, Blonde L. Efficacy and safety of the human glucagon-like peptide-1 analog liraglutide in combination with metformin and thiazolidinedione in patients with type 2 diabetes (LEAD-4 Met+TZD). Diabetes Care. 2009;32(7): 1224-1230. https://doi.org/10.2337/dc08-2124.

[8] Schwartz A V., Chen H, Ambrosius WT, Sood A, Josse RG, Bonds DE, Schnall AM, Vittinghoff E, Bauer DC, Banerji MA, Cohen RM, Hamilton BP, Isakova T, Sellmeyer DE, Simmons DL, Shibli-Rahhal A, Williamson JD, Margolis KL. Effects of TZD Use and Discontinuation on Fracture Rates in ACCORD Bone Study. J Clin Endocrinol Metab. 2015;100(11): 4059-4066. https://doi.org/10.1210/jc.2015-1215.

[9] Chhajed SS, Shinde PE, Kshirsagar SJ, Sangshetti J, Gupta PP, Parab M, Dasgupta D. De-novo design and synthesis of conformationally restricted thiazolidine-2,4-dione analogues: highly selective PPAR-γ agonist in search of anti-diabetic agent. Struct Chem. 2020;31(4): 1375-1385. https://doi.org/10.1007/s11224-020-01500-4.

[10] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7): 476-488. https://doi.org/10.1002/minf.201000061.

[11] Tropsha A, Golbraikh A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr. Pharm. Des. 2007;13(34): 3494-3504. https://doi.org/10.2174/138161207782794257

[12] Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;432(7019): 862-865. https://doi.org/10.1038/nature03197.

[13] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin I I, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: Where have you been? Where are you going to? J Med Chem. 2014;57(12): 4977-5010. https://doi.org/10.1021/jm4004285.

[14] Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. J Mol Graph Model. 2008;26(8): 1315-1326. https://doi.org/10.1016/j.jmgm.2008.01.002.

[15] Wang Z, Chen J, Hong H. Developing QSAR models with defined applicability domains on PPARγ binding affinity using large data sets and machine learning algorithms. Environ Sci Technol. 2021;55(10): 6857-6866. https://doi.org/10.1021/acs.est.0c07040.

[16] Saxena A, Mathur N, Pathak P, Tiwari P, Mathur SK. Machine learning model based on ınsulin resistance metagenes underpins genetic basis of type 2 diabetes. Biomolecules. 2023;13(3): 432. https://doi.org/10.3390/biom13030432.

[17] Lee S, Zhou J, Wong WT, Liu T, Wu WKK, Kei Wong IC, Zhang Q, Tse G. Glycemic and lipid variability for predicting complications and mortality in diabetes mellitus using machine learning. BMC Endocr Disord. 2021;21(1): 94. https://doi.org/10.1186/s12902-021-00751-4.

[18] Kim JY, Han JM, Yun B, Yee J, Gwak HS. Machine learning-based prediction of risk factors for abnormal glycemic control in diabetic cancer patients receiving nutrition support: a case–control study. Hormones. 2023;22(4): 637-645. https://doi.org/10.1007/s42000-023-00492-0.

[19] Chen S, Phuc PT, Nguyen P, Burton W, Lin SJ, Lin WC, Lu CY, Hsu MH, Cheng CT, Hsu JC. A novel prediction model of the risk of pancreatic cancer among diabetes patients using multiple clinical data and machine learning. Cancer Med. 2023;12(19): 19987-19999. https://doi.org/10.1002/cam4.6547.

[20] Yang L, Gabriel N, Hernandez I, Winterstein AG, Guo J. Using machine learning to identify diabetes patients with canagliflozin prescriptions at high-risk of lower extremity amputation using real-world data. Pharmacoepidemiol Drug Saf. 2021;30(5): 644-651. https://doi.org/10.1002/pds.5206.

[21] Yang L, Gabriel N, Hernandez I, Guo S. PDG82 machine learning to ıdentify diabetes patients with canagliflozin prescriptions at high-risk of lower extremity amputation using real-word DATA. Value Heal. 2020;23: S532. https://doi.org/10.1016/j.jval.2020.08.765.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

[22] Yang L, Gabriel N, Hernandez I, Vouri SM, Kimmel SE, Bian J, Guo J. Identifying patients at risk of acute kidney injury among medicare beneficiaries with type 2 diabetes initiating SGLT2 inhibitors:A machine learning approach. Front Pharmacol. 2022;13: 834743. https://doi.org/10.3389/fphar.2022.834743.

[23] Ma J, Theiler J, Perkins S. Accurate on-line support vector regression. Neural Comput. 2003;15(11): 2683-2703. https://doi.org/10.1162/089976603322385117.

[24] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7: 21. https://doi.org/10.3389/fnbot.2013.00021.

[25] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev. 2015;71: 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001.

[26] Rathore SS, Kumar S. A Decision Tree regression based approach for the number of software faults prediction. ACM SIGSOFT Softw Eng Notes. 2016;41(1): 1-6. https://doi.org/10.1145/2853073.2853083.

[27] Bisong E. Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress; 2019: 59-64. https://doi.org/10.1007/978-1-4842-4470-8_7.

[28] Kumar P, Duhan M, Sindhu J, Kadyan K, Saini S, Panihar N. Thiazolidine-4-one clubbed pyrazoles hybrids: Potent α-amylase and α-glucosidase inhibitors with NLO properties. J Heterocycl Chem. 2020;57(4): 1573-1587. https://doi.org/10.1002/jhet.3882.

[29] Khan SA, Ali M, Latif A, Ahmad M, Khan A, Al-Harrasi A. Mercaptobenzimidazole-based 1,3-thaizolidin-4-ones as antidiabetic agents: Synthesis, ın vitro α-glucosidase ınhibition activity, and molecular docking studies. ACS Omega. 2022;7(32): 28041-28051. https://doi.org/10.1021/acsomega.2c01969.

[30] Patil VM, Tilekar KN, Upadhyay NM, Ramaa CS. Synthesis, ın-vitro evaluation and molecular docking study of n-substituted thiazolidinediones as α-glucosidase ınhibitors. ChemistrySelect. 2022;7(1). https://doi.org/10.1002/slct.202103848.

[31] Gummidi L, Kerru N, Ebenezer O, Awolade P, Sanni O, Islam MS, Singh P. Multicomponent reaction for the synthesis of new 1,3,4-thiadiazole-thiazolidine-4-one molecular hybrids as promising antidiabetic agents through α-glucosidase and α-amylase inhibition. Bioorg Chem. 2021;115: 105210. https://doi.org/10.1016/j.bioorg.2021.105210.

[32] Chinthala Y, Kumar Domatti A, Sarfaraz A, Pratap Singh S, Kumar Arigari N, Gupta N, K V N Satya S, Kotesh Kumar J, Khan F, Tiwari AK, Paramjit G. Synthesis, biological evaluation and molecular modeling studies of some novel thiazolidinediones with triazole ring. Eur J Med Chem. 2013;70: 308-314. https://doi.org/10.1016/j.ejmech.2013.10.005.

[33] Bhutani R, Pathak DP, Kapoor G, Husain A, Iqbal MA. Novel hybrids of benzothiazole-1,3,4-oxadiazole-4-thiazolidinone: Synthesis, in silico ADME study, molecular docking and in vivo anti-diabetic assessment. Bioorg Chem. 2019;83: 6-19. https://doi.org/10.1016/j.bioorg.2018.10.025.

[34] Khan T, Lawrence AJ, Azad I, Raza S, Joshi S, Khan AR. Computational drug designing and prediction of ımportant parameters using in silico methods- A review. Curr Comput Aided Drug Des. 2019;15(5): 384-397. https://doi.org/10.2174/1573399815666190326120006.

[35] Consonni V, Todeschini R. Molecular Descriptors. In: Molecular Descriptors for Chemoinformatics; 2010: 29-102. https://doi.org/10.1007/978-1-4020-9783-6_3.

[36] Mauri A, Consonni V, Todeschini R. Molecular Descriptors. In: Handbook of Computational Chemistry. Springer International Publishing; 2017: 2065-2093. https://dx.doi.org/10.1007/978-3-319-27282-5_51.

[37] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7): 1466-1474. https://doi.org/10.1002/jcc.21707.

[38] Randić M. Generalized molecular descriptors. J Math Chem. 1991;7(1): 155-168. https://doi.org/10.1007/BF01200821

[39] Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors Actuators B Chem. 2015;212: 353-363. https://doi.org/10.1016/j.snb.2015.02.025.

[40] Rajput A, Thakur A, Mukhopadhyay A, Kamboj S, Rastogi A, Gautam S, Jassal H, Kumar M. Prediction of repurposed drugs for Coronaviruses using artificial intelligence and machine learning. Comput Struct Biotechnol J. 2021;19: 3133-3148. https://doi.org/10.1016/j.csbj.2021.05.037.

[41] Panahi M, Sadhasivam N, Pourghasemi HR, Rezaie F, Lee S. Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). J Hydrol. 2020;588: 125033. https://doi.org/10.1016/j.jhydrol.2020.125033.

[42] Zhang F, O'Donnell LJ. Support vector regression. In: Machine Learning. Elsevier; 2020: 123-140. https://doi.org/10.1016/B978-0-12-815739-8.00007-9.

[43] Smith PF, Ganesh S, Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J Neurosci Methods. 2013;220(1): 85-91. https://doi.org/10.1016/j.jneumeth.2013.08.024.

[44] Tso GKF, Yau KKW. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy. 2007;32(9): 1761-1768. https://doi.org/10.1016/j.energy.2006.11.010.

[45] Rakhra M, Soniya P, Tanwar D, Singh P, Bordoloi D, Agarwal P, Takkar S, Jairath K, Verma N. Crop price prediction using random forest and decision tree regression:-A review. Mater Today Proc. https://doi.org/10.1016/j.matpr.2021.03.261.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

[46] Hepp T, Schmid M, Gefeller O, Waldmann E, Mayr A. Approaches to regularized regression – A comparison between gradient boosting and the Lasso. Methods Inf Med. 2016;55(05): 422-430. http://dx.doi.org/10.3414/me16-01-0033.

[47] Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1): 1. https://doi.org/10.1186/s13321-017-0256-5.

[48] Nyirandayisabye R, Li H, Dong Q, Hakuzweyezu T, Nkinahamira F. Automatic pavement damage predictions using various machine learning algorithms: Evaluation and comparison. Results Eng. 2022;16: 100657. https://doi.org/10.1016/j.rineng.2022.100657.

[49] Probst P, Wright MN, Boulesteix A. Hyperparameters and tuning strategies for random forest. WIREs Data Min Knowl Discov. 2019;9(3):e1301. https://doi.org/10.1002/widm.1301.

[50] Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol Modell. 2019;406: 109-120. https://doi.org/10.1016/j.ecolmodel.2019.06.002.

[51] Joy TT, Rana S, Gupta S, Venkatesh S. Hyperparameter tuning for big data using Bayesian optimisation. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE; 2016: 2574-2579. https://doi.org/10.1109/ICPR.2016.7900023.

[52] Ito K, Nakano R. Optimizing Support Vector regression hyperparameters based on cross-validation. In: Proceedings of the International Joint Conference on Neural Networks, 2003. Vol 3. IEEE; : 2077-2082. https://doi.org/10.1109/IJCNN.2003.1223728.

[53] Laref R, Losson E, Sava A, Siadat M. On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications. Chemom Intell Lab Syst. 2019;184: 22-27. https://doi.org/10.1016/j.chemolab.2018.11.011.

[54] Aghaaminiha M, Mehrani R, Reza T, Sharma S. Comparison of machine learning methodologies for predicting kinetics of hydrothermal carbonization of selective biomass. Biomass Convers Biorefinery. 2023;13(11): 9855-9864. https://doi.org/10.1007/s13399-021-01858-3.

[55] Santos CE da S, Sampaio RC, Coelho L dos S, Bestard GA, Llanos CH. Multi-objective adaptive differential evolution for SVM/SVR hyperparameters selection. Pattern Recognit. 2021;110: 107649. https://doi.org/10.1016/j.patcog.2020.107649.

[56] Faris, H, Hassonah MA, Al-Zoubi AM, Mirjalili S, Aljarah I. A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture. Neural Computing and Applications. 2018;30: 2355-2369. https://doi.org/10.1007/s00521-016-2818-2.

[57] Hoque KE, Aljamaan H. Impact of hyperparameter tuning on machine learning models in stock price forecasting. IEEE Access. 2021;9: 163815-163830. https://doi.org/10.1109/ACCESS.2021.3134138.

[58] Bormans JP. Systematics of mean resonance spacing and average radiative width from random forest regression.

[59] Dhiyaussalam, Wibowo A, Nugroho FA, Sarwoko EA, Setiawan IMA. Classification of Headache Disorder Using Random Forest Algorithm. In: 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). IEEE; 2020: 1-5. https://doi.org/10.1109/ICICoS51170.2020.9299105.

[60] Abebe M, Shin Y, Noh Y, Lee S, Lee I. Machine learning approaches for ship speed prediction towards energy efficient shipping. Appl Sci. 2020;10(7): 2325. https://doi.org/10.3390/app10072325.

[61] Qi C, Chen Q, Dong X, Zhang Q, Yaseen ZM. Pressure drops of fresh cemented paste backfills through coupled test loop experiments and machine learning techniques. Powder Technol. 2020;361: 748-758. https://doi.org/10.1016/j.powtec.2019.11.046.

[62] Hasanipanah M, Faradonbeh RS, Armaghani DJ, Amnieh HB, Khandelwal M. (2017). Development of a precise model for prediction of blast-induced flyrock using regression tree technique. Environmental Earth Sciences. 2017;76: 1-10. https://doi.org/10.1007/s12665-016-6335-5

[63] Kolan A, Moukthika D, Sreevani KSS, Jayasree H. Click-through rate prediction using decision tree. Proceedings of the Third International Conference on Computational Intelligence and Informatics Advances in Intelligent Systems and Computing 2020: 29-37. https://doi.org/10.1007/978-981-15-1480-7_3.

[64] Li M. Application of CART decision tree combined with PCA algorithm in intrusion detection. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE; 2017: 38-41. https://doi.org/10.1109/ICSESS.2017.8342859.

[65] Li G, Sun Y, Qi C. Machine learning-based constitutive models for cement-grouted coal specimens under shearing. Int J Min Sci Technol. 2021;31(5): 813-823. https://doi.org/10.1016/j.ijmst.2021.08.005.

[66] Charoen-Ung P, Mittrapiyanuruk P. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques. In: 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE; 2018: 1-6. https://doi.org/10.1109/JCSSE.2018.8457391.

[67] Dutta J, Kim YW, Dominic D. Comparison of Gradient Boosting and Extreme Boosting Ensemble Methods for Webpage Classification. In: 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEE; 2020: 77-82. https://doi.org/10.1109/ICRCICN50933.2020.9296176.

[68] Gao R, Liu Z. An ımproved AdaBoost algorithm for hyperparameter optimization. J Phys Conf Ser. 2020;1631(1): 012048. https://doi.org/10.1088/1742-6596/1631/1/012048.

[69] Di Bucchianico A. Coefficient of Determination (R2). In: Encyclopedia of Statistics in Quality and Reliability. Wiley; 2007. https://doi.org/10.1002/9780470061572.eqr173.

Ghosh et al.
Machine Learning based QSAR model for therapeutically active
candidate drugs with Thiazolidinedion scaffold

Journal of Research in Pharmacy

Research Article

[70] Plonsky L, Ghanbar H. Multiple regression in L2 research: A methodological synthesis and guide to ınterpreting R^2 values. Mod Lang J. 2018;102(4): 713-731. https://doi.org/10.1111/modl.12509.

[71] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7(3): 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014.

[72] Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30: 79-82. https://doi.org/10.3354/cr030079.

[73] Wang W, Lu Y. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. IOP Conf Ser Mater Sci Eng. 2018;324: 012049. https://doi.org/10.1088/1757-899X/324/1/012049.