IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 April 2023, accepted 25 April 2023, date of publication 1 May 2023, date of current version 31 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272231

RESEARCH ARTICLE

Prediction of Inhibition Activity of Dihydrofolate Reductase Inhibitors With Multivariate Adaptive Regression Splines

ZANIB QAYYUM[1], TAHIR MEHMOOD[1], AND LAILA A. AL-ESSA[2]

[1]Department of Mathematics, School of Natural Sciences, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Tahir Mehmood (tahime@gmail.com)

This work was supported by the Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Project PNURSP2023R443.

ABSTRACT Dihydrofolate reductase (DHFR) enzyme is a crucial component of cell growth and proliferation in the human body, making it an important target for treating cancer diseases. This study aims to predict the inhibitory activity (pXC50) of dihydrofolate reductase inhibitors in terms of the quantitative structure-activity relationship (QSAR) model. Interpretation of the QSAR model is vital for understanding the physicochemical processes and to assist structural optimisation. Multivariate adaptive regression splines (MARS), a non-parametric technique, is proposed to model the non-linear relationship between the predictor variables and the response variable of a high-dimensional dataset. The dataset used in this research consists of pXC50 activity of 778 DHFR inhibitors. For our study, the data is divided into 80% training set for model building and 20% testing set for model validation. In comparison, the baseline methods deep neural network (DNN) and partial least squares (PLS) are also applied to QSAR modeling. The testing results show that MARS has the best prediction accuracy according to different measures, where RMSE, MAE, MAPE, and RMSPE are 0.96, 0.69, 0.11, and 0.15 respectively. The efficiency of MARS is apparent in its robust interaction of variables, prediction accuracy, and ability to overcome the neural network's black box system. Thus, MARS technique can be considered an excellent tool for modeling QSAR high-dimensional datasets while exploring the non-linear patterns of data.

INDEX TERMS Regression, multivariate adaptive regression splines, neural network, quantitative structure-activity (QSAR), dihydrofolate reductase inhibitors.

I. INTRODUCTION

Dihydrofolate reductase (DHFR) is suggested as a major enzyme involved in the cell replication process, and therefore, it is an interesting target for the treatment of cancer disease. The enzyme DHFR catalyzes the synthesis of dihydrofolate to tetrahydrofolate (THF). DHFR inhibitors weaken THF, ultimately slowing down the DNA synthesis and cell proliferation. Therefore, the chemical compounds that can be used as DHFR inhibitors can be synthesized using this reaction [1].

The quantitative structure-activity relationship model (QSAR) explains the relationships between the biological activities and the structural properties of chemical compounds [2]. The structural properties of chemical compounds are calculated as molecular descriptors using advanced software [3]. QSAR models are built using wide-ranging statistical and machine learning models. In recent years, scientists are showing more interest in using QSAR technology to evaluate the inhibitory activity of new compounds [4].The traditional trial and error approach of testing the biological activity of newly designed compounds through in vivo or in vitro experimental techniques is a time-consuming and expensive process. Therefore, QSAR applications can reduce the costs and failure rate of experiments. QSAR performs as an in-silico tool to prioritize chemicals concerning their biological activities, resulting in reducing the number of candidate chemicals required for testing with in vivo and in vitro experiments [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

The application of this approach has become of significant importance in the fields of biochemistry, computational chemistry, and medicinal chemistry [6].

QSAR's first mathematical equation $\phi(C)$ formulated by Crum-Brown and Fraser in 1868 [7]. In his work, he carried out his experiments on alkaloids and their methyl derivatives and suggested that there is a strong connection between the physiological activity of molecules and their chemical structures. However, it is believed that Hansch et al. in 1962 laid the foundation of modern QSAR [8]. Based on Veldstra study, he proposed the relationship between lipophilicity and biological potency [9]. Since then, many linear methods and non-linear methods are employed in research practice by computational chemists to calculate QSAR models. Linear methods like multiple linear regression (MLR) and partial least squares (PLS) are commonly used for QSAR modeling in literature. For instance, 1D and 2D QSAR models to predict the toxicity of new nitroaromatic compounds and to investigate the relationship between their substituents and toxicity properties can be developed using the PLS regression technique [10]. Multiple linear regression is used to model a collection of HEPT derivatives to predict their inhibitory activity against HIV-1 reverse transcriptase [11]. However, these models are limited in theirability to interpret non-linear relationships between predictor variables and response variables. In the field of computational chemistry, researchers face the challenge of dealing with hundreds of candidate compounds, each with thousands of calculated molecular descriptors due to the availability of advanced molecular generation tools like Dragon 6 which can calculate up to 4885 molecular descriptors. This problem is known as high dimensionality. To tackle this challenge, advanced non-linear methods like artificial neural networks (ANN) are used, which can model non-linear relationships and also serve as a powerful computational tool to deal with high dimensionality. In the literature, ANN has been used in comparison to partial least squares regression (PLS-R) to develop QSAR models for predicting the inhibitory activity of isonicotinamide derivatives that act as inhibitors of Glycogen synthase kinase-3 beta (GSK-3β) [12]. ANN model proved to be the best predictive model as compared to PLS-R. The high-dimensional dataset can increase the computational cost and reduce the accuracy of non-linear methods. Therefore, dimension reduction technique like principal component analysis (PCA) can play a key role. For example, to predict the biological activity of CCR1 antagonists a method comprising of two stages was employed [13].

MARS is a non-parametric regression technique that produces simple and easily interpret able models, making it a useful approach for QSAR modeling. It can map complex non-linear relationships, provides information on statistical significance of independent variables, and performs feature selection. Some of the previous applications of MARS in the field of computational chemistry include building QSAR models to predict the crystallinity property of bent-core compounds [14], to predict the inhibitory activity of pyridine N-oxide derivatives which act as an inhibitor for SARS [15], [16], to predict the inhibitory activity of 1-(3,3-diphenylpropyl)- piperidinyl derivatives [17], to predict the antitumor activity of acridinone derivatives [18], to predict the antiplasmodial activity of artemisinin compounds [19], to predict the retention effect of alkanes in gas chromatography [20], to predict the bioconcentration factor of polychlorinated biphenyls [21], for modeling blood-brain barrier passage [22].

The primary goal of QSAR modeling using different statistical techniques is to produce a QSAR model with good predictive performance that can further be used to predict the inhibitory activity of novel, untested chemical compounds. A good predictive QSAR model not only does prediction but also does feature selection to determine which structural features of a chemical compound are contributing most to its inhibitory effect.

In the current study, three methods are used in comparison to develop a useful QSAR model which can predict the inhibitory activity of dihydrofolate reductase (DHFR) inhibitors. Multivariate adaptive regression splines (MARS) [23], deep neural network (DNN) [24] and partial least squares (PLS) [25], [26] are considered for this purpose. The paper is structured into four sections. Section II includes a short description of the dataset, approach for data splitting and the methodology of MARS and DNN. Section III details the discussion of modeling results. Section IV mentions the conclusion.

II. MATERIALS AND METHODS

A. DATA DESCRIPTION

The dataset is taken from mendeley database [2], [27]. IC50 values of 778 chemical compounds were measured against the dihydrofolate reductase (DHFR) enzyme of homosapiens. Standard fingerprint representation FCFP4 is used to describe the chemical structure of DHFR inhibitors. The chemical structures of these compounds are calculated using RDKit. Total of 1024 bits FCFP4 fingerprint representation are calculated [28]. The dataset includes a Boolean variable that indicates the presence or absence of a molecular substructure. The IC50 value is a measure of the concentration of a drug required to inhibit 50% of the proteins. In this dataset, the IC50 values are represented as floating-point response variables that are normalized by taking the negative logarithm (pXC50). The FCFP4 fingerprint representation used in this study consists of 1024 bits, where each bit is considered as one predictor variable.

B. DATA SPLITTING

The dataset is split randomly into two parts: an 80% training set, which contains 622 samples, and a 20% test set, which contains 156 samples.

C. PERFORMANCE MEASURE CRITERIA

Table 1 outlines the four performance measure criteria and their respective definitions used for this study. These

TABLE 1. Performance measures and their definitions.

Performance metric	Calculation
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(f_i-y_i)^2}{n}}$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - f_i\rvert$
MAPE	$\frac{\sum_{i=1}^{n}\left\lvert\frac{f_i-y_i}{y_i}\right\rvert}{n} \times 100\%$
RMSPE	$\sqrt{\frac{\sum_{i=1}^{n}(f_i-y_i)^2}{n}} \times 100\%$

criteria are used to compare the prediction ability of MARS and DNN. Root mean squared error (RMSE), mean absolute error(MAE), mean absolute percentage error (MAPE), root mean squared percentage error (RMSPE) evaluate the accuracy of the models. Smaller values of these measures represent greater model accuracy.

D. PREDICTING INHIBITORY ACTIVITY

In our current study, we have utilized MARS [23], DNN [24] and PLS [25], [26] for the purpose of modeling and prediction of bio-activities of DHFR inhibitors. The details of these methods are given below.

1) MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

In 1991, Friedman proposed a multivariate, non-parametric regression method known as multivariate adaptive regression splines (MARS) [23]. This technique models the relationship between an input space consisting of predictor variables, $X\ (n \times p)$ and an output space consisting of target variable, $y\ (n \times 1)$. Mathematical equation of MARS model is written as follows:

$$y = f\ (X) + e \tag{1}$$

In the above equation $e\ (n \times 1)$ is representing the residual vector. MARS is an improved version of Classification and Regression Trees $(CART)$ [29]. As MARS is a non-parametric method, it does not make any assumptions regarding the functional relationship between the input variables and the target variable. Using the predictor variables of the given dataset, MARS model drives a collection of coefficients and piece-wise polynomials of power q. MARS model is built by joining piece-wise polynomials smoothly. These piece-wise polynomials are called splines. These splines are fitted in such a way that they separate the independent variables data into different regions. The knot locations are represented by t. The power q of the splines determines their linearity or non-linearity. For a variable x, MARS calculates splines represented by the following equations:

$$[-(x-t)]_+^q = \begin{cases} (t-x)^q, & if\ x < t, \\ 0, & otherwise \end{cases} \tag{2}$$

$$[+(x-t)]_+^q = \begin{cases} (t-x)^q, & if\ x \geq t, \\ 0, & otherwise \end{cases} \tag{3}$$

In the above equations, $q \geq 0$ and it essentially influences the smoothness of the final estimated MARS model. In our study,

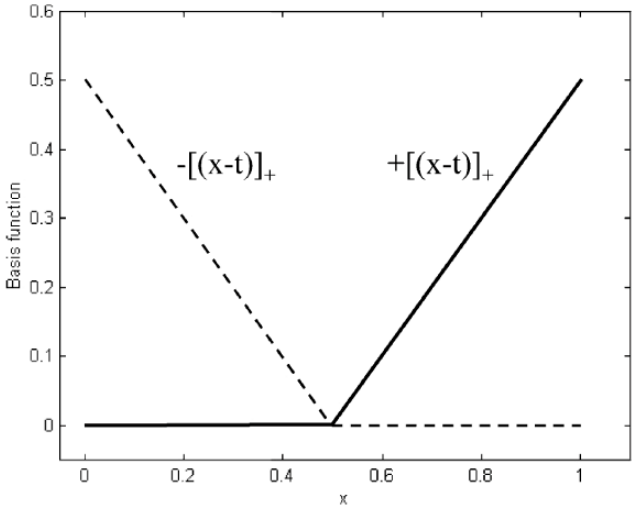


FIGURE 1. Graph representing a pair of splines for variable x; dashed lines showing the left spline $(x < t, -(x-t))$ and solid black line showing the right spline $(x > t, +(x-t))$.

$q = 1$ which means linear splines are measured. Figure 1 shows a pair of splines for a single variable x with $q = 1$ and a knot t. The left spline results in positive values while the right spline remains zero for the values of x variable present on the left side of knot t. For the values of x variable present on the right side of knot t, right spline results in positive values while the left spline remains zero. These splines are also known as basis functions of the variable x. The MARS model of target variable y consisting a set of M basis functions is written as follows:

$$\hat{y} = \hat{f}_M\ (x) = a_o + \sum_{m=1}^{M} a_m B_m\ (x) \tag{4}$$

In the above equation, $\hat{y}$ is the target variable estimated by MARS model, a_o is the intercept term; a constant, $B_m\ (x)$ represents the m_{th} basis function either a single basis function or a product of two or more basis functions whereas a_m is the coefficient of the corresponding m_{th} basis function. The coefficients are estimated by using least squares method.

MARS model in equation 4 is built using a two-step procedure: forward pass and backward pass. The forward pass is completed by adding the predictor variables to the model and optimising the knot positions for each of them. This leads to create two-sided basis functions for each predictor. For the X matrix consisting of n samples and p predictors, there are $n \times p$ pairs of basis functions, represented in equation 2 and equation 3, with knots $x_{ij}\ (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$. Each data point of a predictor variable is considered as a potential knot for the pair of basis functions of that variable. The forward pass produces a complicated and overfitted model which has a poor predictive ability. During the backward pass, the redundant basis functions with the lowest contributions are gradually deleted until the best sub-model with the lowest Generalized Cross Validation (GCV) is produced, improving

the prediction power of MARS model. The GCV formula is written as:

$$GCV\ (M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_M\ \hat{}\ (x_i)\right)^2}{\left(1 - \frac{M + d \times (M-1)/2}{n}\right)^2} \tag{5}$$

where M is the number of basis functions, d is the penalty representing the highest power of the sub-model, n is the number of samples used to build MARS model, and $f_M\ \hat{}\ (x_i)$ represents the target variable estimated by MARS model. The numerator denotes the mean square of residuals of the trained model, and denominator denotes that with an increase in model complexity there is an increase in variance. (M-1)/2 is the number of basis functions knots. The generalized cross-validation technique optimizes number of basis functions and number of knots.

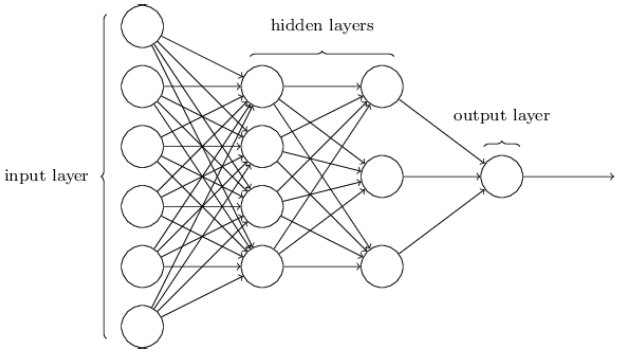
In our study, default value of d parameter is used which is further explained in this article [23]. The MARS model is tuned with two hyperparameters; the maximum interaction level (*degree*) and the maximum number of terms retained in the final model (*nprune*). The degree is restricted from 1 : 4 and nprune is set equal to 50. The ideal combination of these hyperparameters that produces the lowest RMSE for the MARS model is chosen. 10-fold cross Validation is performed on the training data to realize an unbiased RMSE.

The final model comprises of the single variable basis functions and the interaction terms involving multiple basis functions. The relative importance of independent variables, on a scale of 0 to 100, to the fitted MARS model can be estimated using analysis of variance (ANOVA) decomposition procedure [23]. To calculate the score, all the terms involving the variable in consideration is deleted from the model, the reduction in fit is measured. For the most important variable receiving high score, the fitted model has highest reduction in fit.

2) DEEP NEURAL NETWORK (DNN)

Neural network architecture and its processing is inspired by the biological system of the cerebral cortex [30]. A multilayer neural network adopted in this study is shown in Figure II-D2. This neural network consists of an input layer, hidden layer/s and an output layer [24]. The input layer is made up of 1024 neurons whereas the output layer is build from one neuron that demonstrates the prediction of pXC50. The number of hidden layers and the number of nodes within each layer determine the architecture of neural network, which in turn affects how well the network performs. The optimal network architecture is decided by using trial and error method [31], [32].

A neural network tuned by high number of parameters is known as deep neural network (DNN). It models highly non-linear function by connecting multiple layers of meaningful representations that are related by non-linear transformations [33]. A list of parameters used to tune the DNN is as follows:



An example of deep neural network.

- **Activation Function:** Like the nervous system, in the DNN, each neuron has connection with every neuron in the adjacent layer [34]. All the connections attached with a neuron receive a weight. The activation function determines whether a node has sufficient information to send a signal to the next layer. In the present study, the activation function named rectified linear unit function is used [35] as:

$$f\ (x) = \begin{cases} 0, & for\ x < 0, \\ x, & for\ x \geq 0 \end{cases} \tag{6}$$

  The present study uses linear activation function for its output layer which is defined as follows:

$$f(x) = x \tag{7}$$

  The DNN chooses a batch of samples for the forward pass, modify the connection weights between neurons, and predict the output. Loss function and the selected performance metric is used to evaluate the DNN's performance. In the backward pass, DNN scans the layers, computing the gradient of the loss function relative to the weights. The weights are modified in the opposite direction of the gradient, and another batch of samples is selected until the loss function (MSE) and metric (MAE) are minimised. This method is called back-propagation. Two parameters are required to perform back-propagation: objective function and optimization algorithm.
- **Objective Function:** The objective function estimates the error of the predicted output relative to actual output. In the present study, mean-square error (MSE) is used as the loss function. The mathematical formula for the MSE loss function is:

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - f_i)^2 \tag{8}$$

  where y_i and f_i represent the actual response value and the predicted response value for the i_{th} sample respectively. N is the sample size used by DNN.
- **Optimization Algorithm:** The present DNN is trained with root mean square propagation learning algorithm

(RMSprop). The RMSprop technique operates on the principle of maintaining a moving average of the recent mean square gradients of a particular weight [36].

$$\begin{aligned} MeanSquare\,(w,t) &= \gamma \times MeanSquare\,(w,t-1) \\ &\quad + (1-\gamma)\,(\nabla E\,(w))^2 \\ RMSprop: w_{new} &= w_{old} - \frac{\alpha}{\sqrt{MeanSquare\,(w,t)}} \\ &\quad \times triangledownE\,(w_{old}) \end{aligned} \tag{9}$$

where γ has a default value of 0.9, w_{new} is the updated weight, w_{old} is the previous weight, E is the output loss calculated by objective function, α is the learning rate and t is the time step. In the present study, learning rates for different DNN architecures are set at 0.001, 0.002, 0.005 and 0.05.

Epochs and Batch size: The term epoch refers to the number of full forward and backward iterations (forward and backward passes) that the entire training data set makes through the model. The batch size is the number of observations processed by the model prior to weight adjustments. In the present study, the epochs are set as 100, whereas the batch size as set as 32.

One of the problems faced by neural network is that when the model is trained with training data, the loss error is small but when same network is fed with testing data, the loss error is large. This is called overfitting. The problem of overfitting can be adjusted by using two techniques; applying validation split and dropout regularization.

Validation Split: To reduce overfitting, we have tuned our model with validation split parameter. In the present study the validation split is set as 20%. The validation split divides the 80% training dataset into 60% training set and 20% validation set. DNN model is trained for the 60% training set and during the training process, for every epoch, the track of both training and validation loss is kept. This is called internal validation procedure. Ideally, we want the training and validation loss of the DNN model to converge. When the total number of epochs are completed, the training stops and the minimum validation set error weights and biases are reported.

Dropout Regularization: Dropout is another technique to reduce overfitting and improve generalization [37]. It basically sets outputs of a number of neurons in a layer to zero during training, resulting in removing their contributions to the activation of neurons in the following layers during forward pass and the weights of these neurons are not updated during backward pass. This is done to prevent complex co-adaptions amongst neurons [38]. In the present study, dropout rate is set at 40% for third hidden layer, 30% for second hidden layer, 20% for third hidden layer.

Metric: The mathematical formula for MAE is mentioned in Table 1. Similar to loss function track record, MAE of the training set and validation set is also monitored.

TABLE 2. The hyperparameters tuned for the MARS model and the performance measures of the training datasets are listed.

Outputs	Value
Maximum interaction	4
No. of basis functions	50
Pruning method	*cv*
No. of cross-validation folds	10
RMSE	1.02
MAE	0.75
GCV	0.55

E. PARTIAL LEAST SQUARES (PLS)

Partial Least Squares (PLS) [25], [26] is a multivariate regression technique that is widely used in chemometrics and related fields. It is particularly useful when dealing with datasets that have a large number of predictors and a relatively small number of samples. The PLS algorithm can be summarized as follows: Given a dataset with n samples and p predictors, and a response variable Y:

1) Standardize the data to have zero mean and unit variance.
2) Choose the number of latent variables (LVs) to include in the model.
3) Initialize the model by selecting the LV.
4) For each subsequent LV, compute the scores and loadings that maximize the covariance between the predictors and the residuals of the response variable, subject to orthonormality constraints.
5) Compute the weights for each predictor by multiplying the corresponding loadings by the standard deviation of the predictor.
6) Compute the predicted values of Y by taking the dot product of the predictor matrix and the weights matrix.
7) Repeat steps 4-6 until the desired number of LVs has been included in the model.

III. RESULTS AND DISCUSSIONS

To compare results derived from two models, this section first discusses the QSAR model constructed using MARS for predicting the pXC50 inhibitory activity of a new DHFR inhibitor. MARS has two parameters; degree and nprune to tune. These parameters are optimised for each pXC50 value of the training data. In order to improve generalization ability of the model and reduce its overfitting on the training dataset, K-fold cross validation (*cv*) is performed. In our study, K is selected to be 10 and RMSE as the performance metric. The model is tested on the tenth fold of the data after being trained on the first nine folds, and the process is then repeated ten times while switching the test fold. Figure 2 shows the model performance defined by RMSE with different levels of degrees and maximum number of terms. It can be seen that with increasing model complexity, RMSE keeps decreasing. Table 2 summarizes the optimal tuned level of hyperparameters of MARS model and the performance measures obtained for the training samples.

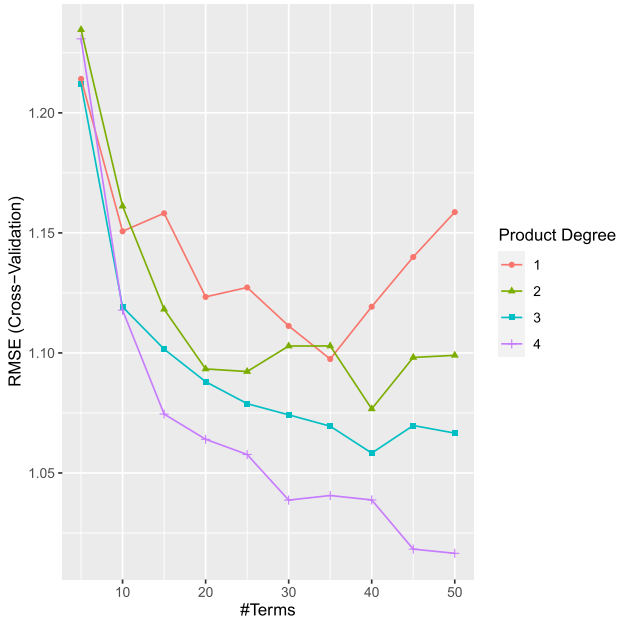


FIGURE 2. MARS model performance defined by RMSECV on y-axis with different number of terms *nprune* on the x-axis. The red color plot represents MARS model is additive, green color plot represents the highest interaction level allowed is 2, blue color plot represents the highest interaction level allowed is 3 and the purple color plot represents the highest interaction level allowed is 4.

Table 3 summarizes the 50 basis functions along with their respective coefficients. In the model, apart from intercept term and 11 basis functions of degree 1, 30 basis functions are of degree 2, 6 basis functions are of degree 3, and 2 basis functions are of degree 4. Interaction among input features reveal that the constructed MARS model is not merely additive, and they play a key role in building an accurate model for capturing the complex and non-linear relationships between the pXC50 activity and a large number of fingerprints of DHFR inhibitors. The coefficient of basis function indicates the magnitude of affect of basis function (input feature) on the output pXC50 activity. A positive sign of an estimated coefficient causes an increase in the pXC50 activity whereas a negative sign results in an opposite effect.

From Table 3, it can be noted that a total of 51 predictors are used in MARS model. The relative importance of predictor variables of the MARS model is assessed through ANOVA decomposition. It is evaluated by deleting the terms relevant to the considered variable from the MARS model, followed by the measurement of the GCV score. High relative importance values represent that the predictor variable has an important contribution to the MARS model and improves the GCV score by a good margin. Table 4 list the 20 most important variables and their corresponding relative importance in terms of percentage. It can be observed that variable b0382 and variable b0988 are contributing as the most important predictors in the MARS model of pXC50, followed by variable b0085, variable b0927 and variable b0840 having the same and high relative importance. These predictor variables play a significant role in determining the pXC50 activity.

TABLE 3. A list of basis functions and their coefficients of the MARS model.

BF	Equation	Coefficient
BF1	1	4.79
BF2	b0017	-0.75
BF3	b0048	1.02
BF4	b0062	0.65
BF5	b0235	2.98
BF6	b0405	-1.20
BF7	b0415	-1.12
BF8	b0679	0.61
BF9	b0726	1.62
BF10	b0736	1.01
BF11	b0895	-1.19
BF12	b0988	1.03
BF13	b0017*b0894	0.76
BF14	b0062*b0153	0.94
BF15	b0062*b0350	0.44
BF16	b0062*b0533	-0.51
BF17	b0062*b0701	2.01
BF18	b0069*b0988	-1.15
BF19	b0196*b0584	1.26
BF20	b0218*b0927	1.44
BF21	b0248*b0927	5.13
BF22	b0305*b0927	-1.75
BF23	b0315*b0988	-0.95
BF24	b0349*b0988	0.72
BF25	b0382*b0988	1.91
BF26	b0387*b0988	3.86
BF27	b0503*b0927	-1.04
BF28	b0514*b0927	-2.41
BF29	b0523*b0988	0.88
BF30	b0559*b0927	-0.91
BF31	b0573*b0988	-0.86
BF32	b0607*b0927	-3.01
BF33	b0654*b0988	1.69
BF34	b0698*b0927	1.80
BF35	b0703*b0988	-0.80
BF36	b0712*b0988	1.93
BF37	b0749*b0988	-1.05
BF38	b0840*b0927	0.90
BF39	b0844*b0988	0.52
BF40	b0853*b0927	2.11
BF41	b0903*b0988	-1.49
BF42	b0925*b0988	0.74
BF43	b0003*b0382*b0988	-1.76
BF44	b0062*b0335*b0533	1.30
BF45	b0085*b0840*b0927	0.86
BF46	b0305*b0599*b0927	2.58
BF47	b0382*b0725*b0988	-1.82
BF48	b0840*b0866*b0927	-2.50
BF49	b0003*b0194*b0382*b0988	1.09
BF50	b0382*b0552*b0725*b0988	2.20

The residual for every fitted response value obtained for training samples is shown in the residual vs fitted graph in Figure 3. As the fitted values increase, the residuals continue to be uniformly distributed demonstrating constant variance that means the model is fulfilling homoscedasticity test. However contrary to linear models, constant variance is less significant for the MARS model. The red curve in the graph is called lowess (locally weighted scatter plot smoothing) fit. Lowess fit represents the mean of the residuals at the fitted values. From the graph it can be seen that mean of the residuals is almost zero and exhibit no deviation at low or high

TABLE 4. A list of 20 most important predictor variables of the MARS model and their relative importance in terms of percentage based on generalized cross-validation (GCV) score.

Variable name	Relative importance (%)
b0382	100.00
b0988	100.00
b0085	89.47
b0927	89.47
b0840	89.47
b0003	83.10
b0017	77.53
b0062	72.46
b0903	69.69
b0895	64.64
b0305	61.89
b0533	59.05
b0701	57.36
b0218	55.26
b0725	53.23
b0349	51.31
b0552	49.18
b0315	47.25
b0248	45.35
b0866	45.53

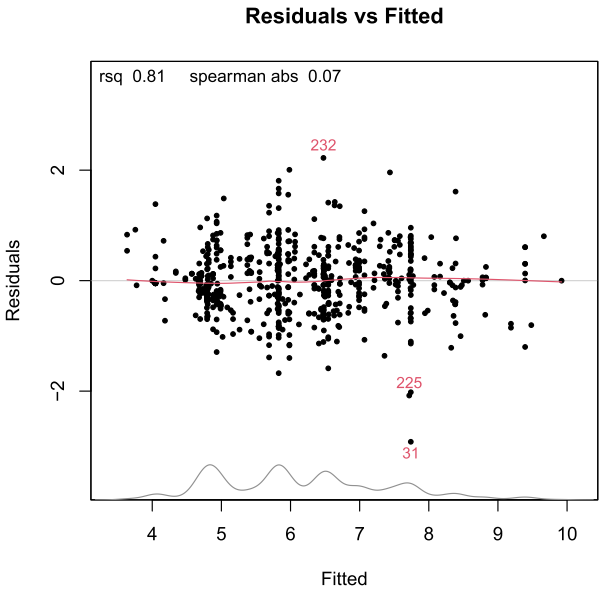


FIGURE 3. A residual versus fitted plot of the MARS model for training data.

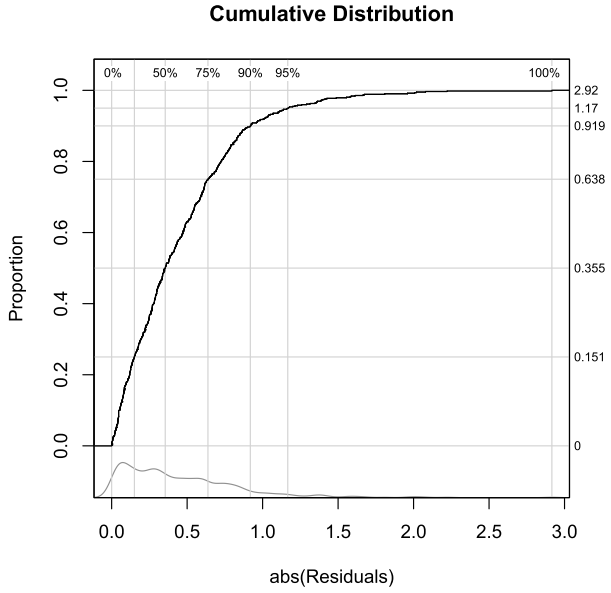


FIGURE 4. A plot representing the cumulative distribution of absolute values of residuals of MARS model.

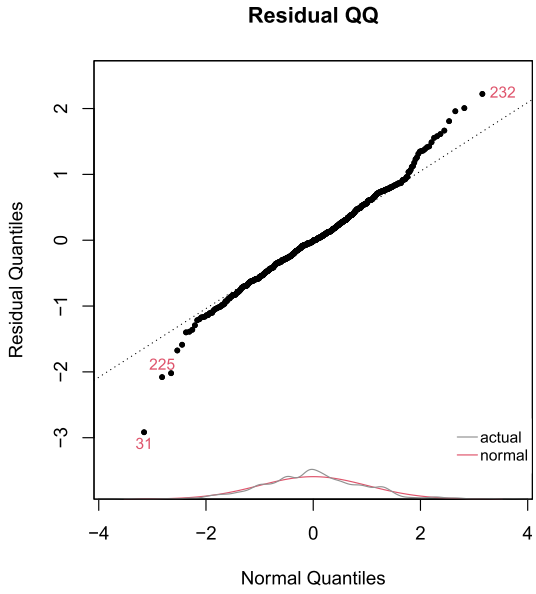


FIGURE 5. Quantile-Quantile (Q-Q) plot of the MARS model.

fitted values. The graph also shows some instances of outliers marked as 31,225 and 232.

The cumulative distribution of the absolute values of residuals is displayed in Figure 4. The ideal graph starts at 0 and quickly rise to 1. The distribution of absolute residual values is represented by the grey coloured curve. The grey distribution curve translates into a black cumulative distribution curve which is nearly S-shaped. The median of absolute values of residuals is 0.335. This means that for the training model, 95% of the times the predicted value is within 1.17 units of the observed value.

The distribution of residuals is compared to normal distribution using Q-Q plot in Figure 5. From the plot it can be seen that residuals are plotted on the black dotted line indicating the residuals follow normal distribution. The normal distribution curve marked as normal meets the distribution of residuals curve marked as actual. The property of normality is not important for MARS model, but it is helpful to detect outliers. The plot does not show any divergence of residuals.

In the next section, QSAR model developed using deep neural network (DNN) and partial least squares (PLS) is discussed. Table 5 summarizes some of the best trial and error experimental results. Results from three neural network architectures optimised with chosen hyperparameters are obtained. As networks with lower learning rates provide

TABLE 5. The hyperparameters optimized for the DNN model and the performance measures of the training and validation datasets are listed.

Network architecture	Learning rate	Training MSE	Validated MSE	Training MAE	Validated MAE
1024-5-1	0.001	0.23	2.73	0.34	1.23
	0.005	0.23	5.53	0.36	1.74
	0.002	0.13	3.50	0.25	1.41
1024-10-5-1	0.001	0.11	3.37	0.24	1.36
	0.005	0.13	3.98	0.28	1.57
	0.002	0.08	2.27	0.21	1.13
1024-100-50-20-1	0.05	1.43	1.23	0.93	0.89

superior results [39], learning rates of 0.001, 0.002, 0.005, and 0.05 are tested during the training process of DNN. All the network architectures are trained for 100 epochs using MSE and MAE as performance measures for the evaluation of training and internal validation. The optimal network architecture has the lowest validated MSE and MAE. The results demonstrate that for single hidden layer neural network $\{1024-5-1\}$, the change of learning rate causes a change in model validation performance. The lowest validation MSE and MAE for $\{1024-5-1\}$ network is achieved at learning rate of 0.001. However, it can be seen there is relatively huge difference between training and validation performance measures due to overfitting of the $\{1024-5-1\}$ network on the training datasets. A new hidden layer with 10 neurons is added to the neural network corresponding to $\{1024-10-5-1\}$ in the Table 5. Results show that the training and validation performances of $\{1024-10-5-1\}$ network have not much improved, and the DNN model still displays overfitting. In the third experiment, a new hidden layer is added to the network and number of neurons are also increased corresponding to $\{1024-100-50-20-1\}$ network in Table 5. The learning rate is increased to 0.05. Furthermore, the network is tested with the application of dropout regularization. In the network, first hidden layer is regularized with a dropout of 40%, second hidden layer with 30%, and third hidden layer with 20%. As a result, the training and validation performances have improved significantly and overfitting has reduced to be negligible. Hence, it can be observed the network architecture $\{1024-100-50-20-1\}$ optimised with a learning rate of 0.05 and a dropout regularization, has the lowest validated MSE and MAE and is considered to be the best network setup for the DNN model for the prediction of pXC50 activity. The experimentation process also shows the importance of dropout regularisation on the performance of DNN. To analyse the convergence properties, training and validated MSE and MAE throughout the training process of the $\{1024-100-50-20-1\}$ network with the learning rate of 0.05 is shown in Figure 6. It is simple to see the good convergence properties of the $\{1024-100-50-20-1\}$ neural network model.

In this section, prediction ability of MARS and DNN models is evaluated using testing data. Table 6 compares the prediction results derived from MARS model and optimal DNN model. Comparison of RMSE, MAE, MAPE and

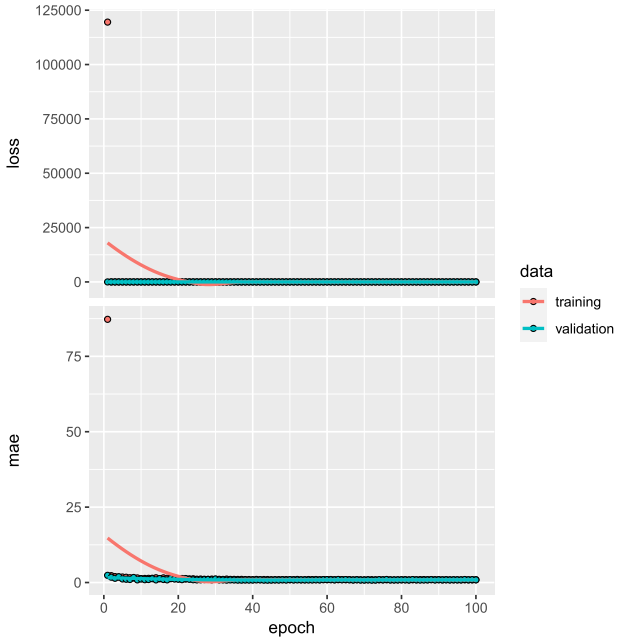


FIGURE 6. Plots showing MSE and MAE performances of $\{1024-100-50-20-1\}$ network for training and validation sets; model parameters: activation function (ReLU), optimization algorithm (RMSprop), learning rate (0.001), dropout rate (40% for first hidden layer, 30% for second hidden layer, 20% for third hidden layer), batch size (32), epoch (100).

RMSPE values shows that DNN shows less accuracy than MARS however both models indicate excellent prediction results. Figure 7 plots the residuals for predicted values of pXC50 using MARS model. One can also conclude that MARS provides a remarkably accurate estimate of the predicted pXC50 activity. Figure 8 plots the residuals for predicted values of pXC50 obtained using optimal DNN model. Comparison of Figure 7 and 8 shows that residuals obtained for DNN model are large than those of the MARS model. Peaks are found for the DNN model at lowest and middle range of the predicted pXC50 activity. This indicates that DNN model is less robust and predictions based on DNN model are less reliable as compared to MARS model. Based on the information given, we can deduce that Partial Least Squares (PLS) has been optimized with 6 components, which means it has potentially improved interpretability due to a reduced feature space. PLS performs better than Deep Neural Networks (DNN) in prediction, possibly because it handles multicollinearity and small sample sizes better than DNN. However, PLS performs worse than Multivariate Adaptive Regression Splines (MARS), which implies that MARS may be better suited for capturing complex nonlinear relationships in the data. It's important to keep in mind that the performance of these algorithms may vary depending on the specific dataset and problem, so it's always recommended to try multiple algorithms and compare their performances.

A. COMPUTATIONS

For computations, modeling and figures of MARS R software and for DNN RStudio is used. For MARS R packages 'earth,'

TABLE 6. Prediction comparison of MARS, DNN and PLS models on testing data.

Methods	RMSE	MAE	MAPE	RMSPE
MARS	0.96	0.69	0.11	0.15
DNN	1.16	0.96	0.17	0.20
PLS	1.37	0.84	0.19	0.17

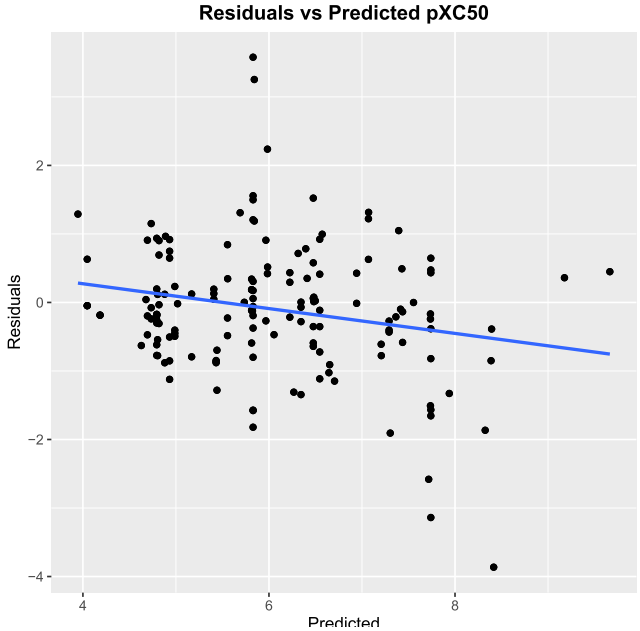


FIGURE 7. A residual plot of predicted values of pXC50 using MARS model.

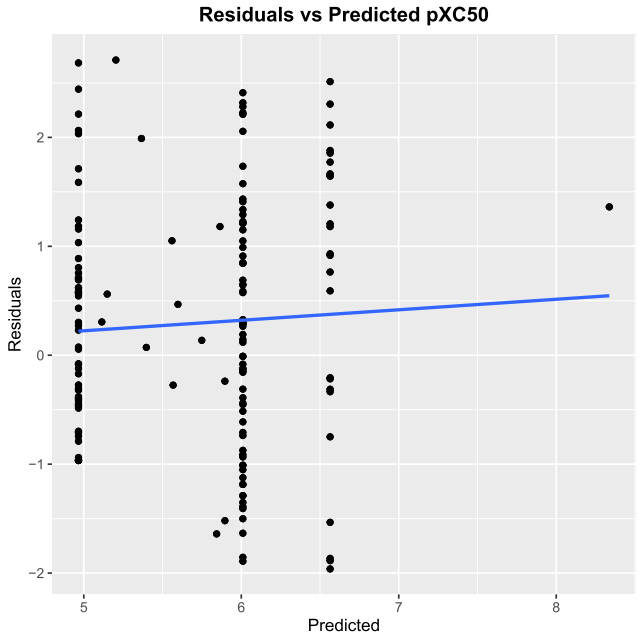


FIGURE 8. A residual plot of predicted values of pXC50 using optimal DNN model.

'caret,' 'vip,' 'tidyverse,' for DNN R packages 'keras,' 'tensorflow,' 'dplyr,' 'magrittr' 'ggplot2,' for PLS 'plsr' and for performance measures R packages 'Metrics' and 'MLmetrics' are used.

IV. CONCLUSION

The research has detailed the steps of constructing the MARS, DNN and PLS models for QSAR study of DHFR inhibitors. Optimal MARS model is identified with 50 BFs comprising of degree 2, 3 and 4. A five layer DNN architecture $\{1024 - 100 - 50 - 20 - 1\}$ is identified as the optimal neural network. Similarly, PLS is optimized with 6 components. According to the prediction measures of RMSE, MAE, MAPE, and RMSPE, MARS proved to be the best modeling technique.

MARS technique is exceptional in the way that it is computationally efficient as compared to neural network. Neural network is criticized for the long training process required to configure the best network. In addition, MARS algorithm is able to compute importance of each variable through ANOVA decomposition procedure. The interpretative property of the MARS model make this technique useful for chemists to optimise structures of chemical compounds.

ACKNOWLEDGEMENT

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R443), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] B. S. Askari and M. Krajinovic, "Dihydrofolate reductase gene variations in susceptibility to disease and treatment outcomes," *Current Genomics*, vol. 11, no. 8, pp. 578–583, Dec. 2010.

[2] I. Olier, N. Sadawi, G. R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova, and R. D. King, "Meta-QSAR: A large-scale application of meta-learning to drug design and discovery," *Mach. Learn.*, vol. 107, no. 1, pp. 285–311, Jan. 2018.

[3] V. Mandlik, P. R. Bejugam, and S. Singh, "Application of artificial neural networks in modern drug discovery," in *Artificial Neural Network for Drug Design, Delivery and Disposition*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 123–139.

[4] K. Z. Myint and X.-Q. Xie, "Recent advances in fragment-based QSAR and multi-dimensional QSAR methods," *Int. J. Mol. Sci.*, vol. 11, no. 10, pp. 3846–3866, Oct. 2010.

[5] J. Hemmerich and G. F. Ecker, "In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways," *Wiley Interdiscipl. Rev., Comput. Mol. Sci.*, vol. 10, no. 4, p. e1475, 2020.

[6] M. A. Elaziz, Y. S. Moemen, A. E. Hassanien, and S. Xiong, "Quantitative structure-activity relationship model for HCVNS5B inhibitors based on an antlion optimizer-adaptive neuro-fuzzy inference system," *Sci. Rep.*, vol. 8, no. 1, pp. 1–17, Jan. 2018.

[7] D. F. Larder, "Alexander crum Brown and his doctoral thesis of 1861," *Ambix*, vol. 14, no. 2, pp. 112–132, Jun. 1967.

[8] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, "Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients," *Nature*, vol. 194, no. 4824, pp. 178–180, Apr. 1962.

[9] H. Veldstra, "The relation of chemical structure to bio-logical activity in growth substances," *Annu. Rev. Plant Physiol.*, vol. 4, no. 1, pp. 151–198, Jun. 1953.

[10] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, L. Gorb, M. Qasim, and J. Leszczynski, "The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study," *J. Computer-Aided Mol. Design*, vol. 22, no. 10, pp. 747–759, Oct. 2008.

[11] J. M. Luco and F. H. Ferretti, "QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives," *J. Chem. Inf. Comput. Sci.*, vol. 37, no. 2, pp. 392–401, Mar. 1997.

[12] A. El Aissouq, H. Toufik, F. Lamchouri, M. Stitou, and A. Ouammou, "QSAR study of isonicotinamides derivatives as Alzheimr's disease inhibitors using PLS-R and ANN methods," in *Proc. Int. Conf. Intell. Syst. Adv. Comput. Sci. (ISACS)*, Dec. 2019, pp. 1–7.
[13] M. Shahlaei, A. Fassihi, and L. Saghaie, "Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: A comparative study," *Eur. J. Medicinal Chem.*, vol. 45, no. 4, pp. 1572–1582, Apr. 2010.
[14] J. Antanasijević, D. Antanasijević, V. Pocajt, N. Trišović, and K. Fodor-Csorba, "A QSPR study on the liquid crystallinity of five-ring bent-core molecules using decision trees, Mars and artificial neural networks," *RSC Adv.*, vol. 6, no. 22, pp. 18452–18464, 2016.
[15] M. Jalali-Heravi, M. Asadollahi-Baboli, and A. Mani-Varnosfaderani, "Shuffling multivariate adaptive regression splines and adaptive neuro-fuzzy inference system as tools for QSAR study of SARS inhibitors," *J. Pharmaceutical Biomed. Anal.*, vol. 50, no. 5, pp. 853–860, Dec. 2009.
[16] Q.-S. Xu, "Multivariate adaptive regression splines—Studies of HIV reverse transcriptase inhibitors," *Chemometric Intell. Lab. Syst.*, vol. 72, no. 1, pp. 27–34, 2004.
[17] M. Jalali-Heravi and A. Mani-Varnosfaderani, "QSAR modeling of 1-(3,3-diphenylpropyl)-piperidinyl amides as CCR5 modulators using multivariate adaptive regression spline and Bayesian regularized genetic neural networks," *QSAR Combinat. Sci.*, vol. 28, no. 9, pp. 946–958, Sep. 2009.
[18] M. Koba and T. Bazek, "The evaluation of multivariate adaptive regression splines for the prediction of antitumor activity of acridinone derivatives," *Medicinal Chem.*, vol. 9, no. 8, pp. 1041–1050, Oct. 2013.
[19] V. Nguyen-Cong, G. Van Dang, and B. Rode, "Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives," *Eur. J. Medicinal Chem.*, vol. 31, no. 10, pp. 797–803, Jan. 1996.
[20] Q.-S. Xu, D. L. Massart, Y.-Z. Liang, and K.-T. Fang, "Two-step multivariate adaptive regression splines for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors," *J. Chromatography A*, vol. 998, nos. 1–2, pp. 155–167, May 2003.
[21] K. Zarei and Z. Salehabadi, "The shuffling multivariate adaptive regression splines and adaptive neuro-fuzzy inference system as tools for QSPR study bioconcentration factors of polychlorinated biphenyls (PCBs)," *Structural Chem.*, vol. 23, no. 6, pp. 1801–1807, Dec. 2012.
[22] E. Deconinck, M. Zhang, F. Petitet, E. Dubus, I. Ijjaali, D. Coomans, and Y. V. Heyden, "Boosted regression trees, multivariate adaptive regression splines and their two-step combinations with multiple linear regression or partial least squares to predict blood–brain barrier passage: A case study," *Analytica Chim. acta*, vol. 609, no. 1, pp. 13–23, 2008.
[23] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, no. 1, pp. 1–67, Mar. 1991.
[24] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiological Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.
[25] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 62–69, Aug. 2012.
[26] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Anal. Chim. Acta*, vol. 185, pp. 1–17, 1986.
[27] I. Olier. (2020). *Qsar Datasets—Meta-Qsar*. Mendeley Data. [Online]. Available: https://doi.org/10.17632/spwgrcnjdg.1
[28] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010.
[29] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[30] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—A new frontier in artificial intelligence research [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.
[31] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin, "Backpropagation: The basic theory," in *Backpropagation: Theory, Architectures and Applications*. New York, NY, USA: Taylor & Francis, 1995, pp. 1–34.
[32] S. Haykin, *Neural Networks, a Comprehensive Foundation*, vol. 7458. Upper Saddle River, NJ, USA: Prentice-Hall, pp. 161–175, 1999.
[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Sep. 2015.
[34] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.
[35] A. Fred Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.
[36] G. Hinton, N. Srivastava, and K. Swersky. (2012). *Neural Networks for Machine Learning-Lecture 6E-Rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude*. [Online]. Available: https://www.cs.toronto.edu/tijmen/csc321/slides/lectureslideslec6.pdf
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014, Accessed: Mar. 23, 2020.
[38] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.
[39] Y. Chauvin and D. E. Rumelhart, *Backpropagation: Theory, Architectures, and Applications*. London, U.K.: Psychology Press, 2013.

ZANIB QAYYUM received the M.Phil. degree in statistics from the School of Natural Sciences, National University of Sciences and Technology, Islamabad, Pakistan, in 2022. Her research interests include high-dimensional data, machine learning, and bio-statistics.

TAHIR MEHMOOD received the Ph.D. degree in statistics from the Norwegian University of Life Science (NMBU), Norway, in 2012. He is currently a Professor of statistics with the School of Natural Science (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interests include multivariate statistics, statistical learning, classification, clustering, variable selection, and the application of these methods/algorithm covers chemometrics, environmetrics, and public health.

LAILA A. AL-ESSA received the Ph.D. degree from Princess Nora bint Abdulrahman University, Saudi Arabia. She is currently an Assistant Professor of mathematics with the Department of Mathematical Sciences, College of Science, Princess Noura bint Abdulrahman University. Her research interests include reliability estimation, probability distributions, and ordinal statistics.