Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:73

https://doi.org/10.28919/cmbn/7549

ISSN: 2052-2541

EVALUATION QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) USING ENSEMBLE LEARNING METHODS ON ACETYLCHOLINESTERASE INHIBITORS FOR ALZHEIMER'S DISEASE

ALHADI BUSTAMAM[1,2,*], MUSHLIHA[1,2], ARRY YANUAR[3], PRASNURZAKI ANKI[1,2], ADAWIYAH ULFA[1]

[1]Department of Mathematics, Universitas Indonesia, Depok, Indonesia

[2]Data Science Centre, Universitas Indonesia, Depok, Indonesia

[3]Department of Pharmacy, Faculty of Pharmacy, Universitas Indonesia, Depok, Indonesia

Abstract: Acetylcholinesterase inhibitors (AChEI) are among the most potential drug molecules for treating Alzheimer's disease and effectively treating its symptoms. Quantitative Structure and Activity Relationship (QSAR) is a computational modeling method to determine the relationship between the structural properties of chemical compounds and biological activities. This study used a classification QSAR model to predict the active and inactive molecules in AChEI. There were 3809 molecules of compounds in the preprocessing stage consisting of 2215 molecules of active compounds and 1594 molecules of inactive compounds. The compound molecules in SMILES were extracted into the fingerprint using the ECFP and FCFP method with diameters of 4 and 6. In this study, the ensemble learning methods used to build the classification QSAR model were voting, averaging, and stacking. The results showed that the ensemble learning method had a better performance than using only one base model. The classification QSAR model with base model obtained an accuracy of 92%, a sensitivity of 89.97%, a specificity of

*Corresponding author

E-mail address: alhadi@sci.ui.ac.id

Received June 13, 2022

93%, and an MCC of 83%. The comparison, the ensemble learning method with the stacking technique obtained an accuracy of 93%, a sensitivity of 92%, a specificity of 94%, and an MCC of 86%.

Keywords: classification; fingerprint; potential drug; SMILE; structural properties.

2010 AMS Subject Classification: 92C50.

1. INTRODUCTION

Alzheimer's disease is characterized by neurodegenerative problems that worsen over time. Memory and cognitive decline increased impairment in daily activities, and a wide range of neuropsychiatric symptoms and behavioral abnormalities are among the symptoms [1]. The prevalence of Alzheimer's disease in people over 60 years is 40.2 per 1000, while the incidence proportion is 34.1 per 1000 [2]. These figures indicate that over 45 million people worldwide are affected by symptoms. Furthermore, until at least 2050, this incidence is predicted to double every 20 years [3]. Given the high prevalence and risk caused by Alzheimer's disease, proper treatment is needed to deal with this problem.

Acetylcholinesterase inhibitors (AChEIs) are among the most potent drug molecules against Alzheimer's disease and effectively treat its symptoms. However, some synthetic acetylcholinesterase inhibitor drugs such as physostigmine donepezil or tacrine, galantamine, and rivastigmine are known to have side effects such as hepatotoxicity and gastrointestinal disturbances [4]. Based on the above, safe Alzheimer's disease drugs using AChEIs must be developed to minimize side effects.

In the early stages of drug development and design, there are often millions of potential therapeutic molecules under consideration. It is critical to anticipate drug candidate action early by utilizing computational (in silico) methodologies to save time and costs. The activity of biological compounds is predictable using QSAR. QSAR is a method for developing mathematical models that aid in understanding the relationship between the chemical structure of molecular

compounds and their biological activity. [5]. QSAR can be implemented with machine learning methods developed for drug discovery [6]. Thus, QSAR using machine learning can help the drug discovery process more effectively and efficiently.

QSAR assumes that compounds with similar structural properties will have similar activities. A key aspect of QSAR is molecular descriptors as numerical representations of chemical structures. QSAR predicts molecular activity using molecular descriptors calculated by molecular graphs, quantum chemical descriptors, and Simplified Molecular-Input Line-Entry System (SMILES) [7]. SMILES is a widely used representation of molecular structure with symbol sequences for QSAR analysis [8]. With molecular descriptors derived from multiple algorithms, molecules can be uniquely identified from chemical databases. SMILES can be used to extract molecular descriptors from compound molecules and turn them into molecular fingerprints (MF). Each compound molecule has a unique vector, namely a fingerprint representing the MF [9]. A fingerprint is a binary representation of the structure and properties of a molecule. The bit-string encodes whether a property is present (1) or absent (0), which can be a chemical structure or a fragment [10]. A fingerprint seems to be a molecular descriptor that aggregates the presence or absence of different molecular substructures inside a molecule into a single molecule. Methods of describing molecules transformed into bit vectors, such as the Extended-Connectivity Fingerprints (ECFP), Functional Class Fingerprints (FCFP), and Molecular ACCess System (MACCS), can be used to classify fingerprints [11]. Therefore, selecting the correct fingerprint can affect the presentation of molecules used in building the QSAR model.

There have been developments in QSAR classification using machine learning [12]. One built a QSAR model using the Decision Tree (DT), Support Vector Machine (SVM), and Neural Network (NN) for DPP-IV inhibitors. These are inhibitors for type 2 Diabetes Mellitus with the best SVM model with performance specificity, sensitivity, accuracy, and Matthews Correlation Coefficient (MCC) is 0.774, 0.826, 0.803, and 0.604, respectively. Other studies [13] built a classification QSAR model to predict the activity of Acetylcholinesterase inhibitors (AChE).

AChE inhibitors treat Alzheimer's disease using Deep Neural Networks (DNN) and Multi-Layer Perceptron (MLP) models with the highest model accuracy, owning DNN by 84%. Classification QSAR performance can improve with ensemble learning methods that combine several models.

Each machine learning model has its advantages and disadvantages. Various model tests have been carried out to obtain the best model, which can be concluded as the most suitable model for a problem in research [14]. Here are some advantages of machine learning models. MLP has several advantages compared to other classification models, namely adaptive models, generating the necessary decision functions directly through training, working with insufficient knowledge, having distributed memory, fault tolerance, and being universal models [15]. The advantages of kNN are simple, do not require data assumptions, high accuracy, and easy to implement [16], and are very effective in predictive performance [17]. LR has a strong reputation as one of the most successful classification tools, with applications spanning machine learning, data mining, pattern recognition, and medical science to statistics. [18]. LR is easier to implement, and interpret, makes no assumptions about the distribution of classes in the feature space, and is very efficient to train [19]. Based on these advantages, this research uses MLP, kNN, and LR as base learners. It then uses the ensemble method to combine the three models to build a classification QSAR model to produce better performance.

This study aimed to build a classification QSAR model for Acetylcholinesterase inhibitors (AChEIs) as drug molecules for treating Alzheimer's disease using the Ensemble Learning method and evaluated the model's performance based on accuracy, sensitivity, specificity, and MCC. The novelty of this study was using the Ensemble Learning method by combining the MLP, kNN, and Logistic Regression (LR) models with voting, averaging and stacking techniques to build a classification QSAR model. AChEI molecular data were obtained from the ChEMBL database site (www.ebi.ac.uk/chembl) with ID CHEMBL220. IC50 measured AChEIs molecular activity with MF using the ECFP (diameter 4 and 6) and FCFP (diameter 4 and 6) methods.

2. Discussion of the Models and Methods

2.1. Inhibitor Acetylcholinesterase

Alzheimer's disease is a slow-progressing neurodegenerative condition and deadly brain disease that affects about 5–10% of the population over 65 [20]. Alzheimer's disease is characterized by a substantial loss of memory and other intellectual faculties that makes daily life difficult. The loss of cholinergic neurons in the brain and a decrease in acetylcholinesterase are linked to this disease (AChE). AChE inhibitors in the brain are the key therapeutic targets in Alzheimer's disease treatment efforts. Cholinesterase inhibitors reduce damage while inhibiting AChE activity and maintaining AChE levels. As a result, AChE boosts cholinergic neurotransmission in the forebrain, making up for the loss of brain cell function [21].

Alzheimer's disease is linked to decreased acetylcholine (ACh) levels and the death of cholinergic neurons in the brain. ACh was the first neurotransmitter discovered, and it transmits nerve signals throughout the autonomic nervous system, including neuromuscular junctions and synapses. In the autonomic nervous system, Ach regulates neurotransmission signals between preganglionic sympathetic and parasympathetic neurons [4]. ACh is also in charge of muscular activation, which includes the digestive tract muscles. The loss of ACh function has ramifications for Alzheimer's disease development. Normal neurotransmission is inhibited by acetylcholinesterase (AChE), an enzyme that converts acetate and choline from the neurotransmitter Ach. According to the cholinergic theory, blocking AChE could be a potential therapy option for Alzheimer's disease symptoms. Therefore, one of the essential objectives in treating Alzheimer's disease is AChE [22].

2.2. Quantitative Structure-Activity Relationship (QSAR)

QSAR is a technique for interdisciplinary compound investigation that includes chemistry, physics, biology, and toxicity components. QSAR is a method for formally establishing mathematical relationships between chemical properties and activity manifestations of structurally comparable substances. All techniques are defined based on robust mathematical algorithms and

provide a sound basis for building predictive correlation models. The QSAR technique, in addition to giving mathematical correlations, also allows the investigation of chemical properties stored in the descriptors [5]. The QSAR technique's basic formula can be calculated numerically represented as follows:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n \quad (1)$$

Where Y is the dependent variable that represents the modeled response, namely activity, while $X_1, X_2, X_3, \ldots, X_n$ is the independent variable that shows different structural features or physicochemical properties in the form of descriptors and $a_1, a_2, a_3, \ldots, a_n$ is the contribution of each descriptor to the response where $a_0$is a constant. The main objectives of QSAR studies are predicting compounds' biological activity, optimizing, designing the active ingredients of new compounds, predicting risk and toxicity assessments, modeling pharmacokinetic and pharmacodynamic profiles of new chemical entities, and finding compounds with the desired biological activity by filtering chemical databases or virtual libraries [5].

2.3. Multi-layer Perceptron (MLP)

MLP is one type of Feedforward Neural Network (FNN) that contains at least three layers: input, hidden, and output. MLP uses the back propagation technique for learning. Therefore, MLP comprises three layers: an input layer of neurons that act as receivers, one or more hidden layers of neurons that calculate data and iterate, and an output layer that forecasts output [25].

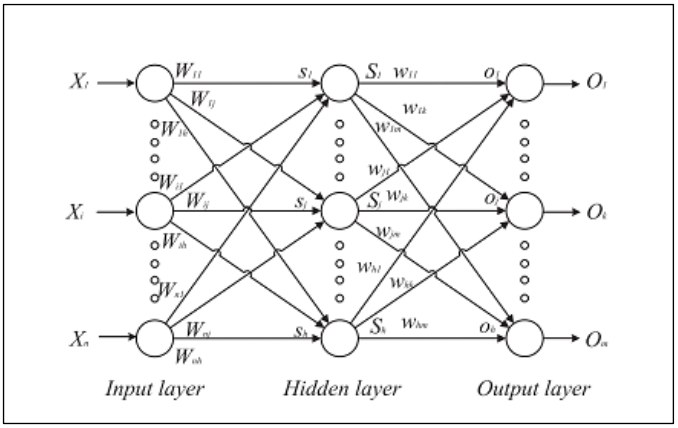


FIGURE 1. The basic structure of MLP [26]

The input nodes number is n, the number of the hidden node is h, and the output nodes number is m in Figure 1, displaying an MLP with three layers. Because MLP is an FNN, there is a one-way link between the nodes. The output of the MLP is calculated as follows:

$$s_j = \sum_{i=1}^{n}(W_{ij}.X_i) + b_j, \quad j = 1, 2, 3, \dots, h \tag{2}$$

Where n is the number of input nodes, W_{ij} indicates the connection weight from the i[th] node in the input layer to the j[th] node in the hidden layer, $b_j$is the bias (threshold) of the j[th] hidden node, and X_i indicates the i[th] input.

The output of each hidden node is calculated as follows:

$$S_j = sigmoid\,(s_j) = \sigma(s_j) = \frac{1}{(1+exp(-s_j))}, \quad j = 1,2,3, \dots, h \tag{3}$$

After calculating the output of the hidden node, the final output is defined as follows:

$$o_k = \sum_{j=1}^{h}(W_{jk}.S_j) + b'_k, \quad k = 1,2,3, \dots, m \tag{4}$$

$$O_k = sigmoid(o_k) = \sigma(o_k) = \frac{1}{(1+exp(-o_k))}, \quad k = 1,2,3, \dots, m \tag{5}$$

Where W_{jk} is the weight of the connection from the j[th] hidden node to the k[th] output node, $b'_k$is the kth output node's bias (threshold), connection weight and bias are the most significant aspects of MLP. The final value of the output is determined by the weights and biases, as shown in the preceding equation. MLP training identifies the best weights and biases for a given input to produce the desired output [26].

2.4. K-Nearest Neighbor (kNN)

K-Nearest Neighbor (kNN or KNN) is a classification method that uses the object's nearest neighboring learning data to classify it. The kNN classifier depicts the k-nearest neighbor's classifier [27]. The primary determining factor in kNN is the number of neighbors, and k is the number of closest neighbors.

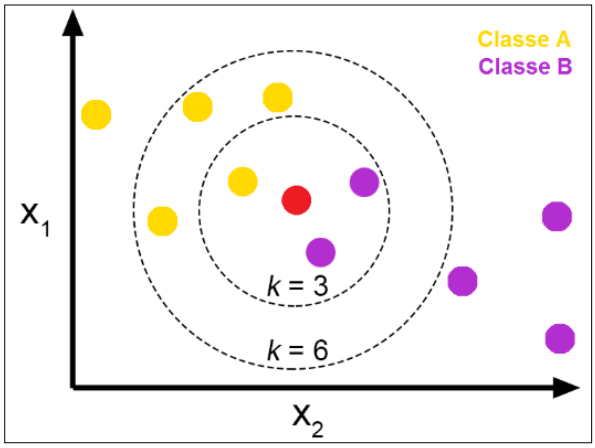


FIGURE 2. Illustration of K Nearest Neighbor in two classes [26]

The kNN algorithm assumes that all examples fit a point in an n-dimensional space. The nearest neighbor of an instance is defined according to the standard Euclidean distance. Let the eigenvectors of x be:

$$AX = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \quad (6)$$

Where $a_r(x)$ represents the value of the r[th] attribute of the x. instance. the distance between the two instances x_i and x_j is defined as $d(x_i,x_j)$ where:

$$d(x_i,x_j) = \sqrt{\sum_{r=1}^{n}\left(a_r(x_i) - a_r(x_j)\right)^2} \quad (7)$$

In nearest neighbor learning, the discrete object classification function is $f: R^n \rightarrow V$, where V is a finite set $\{v_1, v_2, v_3, \ldots, v_n\}$ vs., i.e., the set of distinct categories. The nearest neighbor k value selection is based on the amount and degree of dispersion in each sample type, and different k values can be generated and selected for various applications. Generally, an object is influenced by its neighbors. The closer the thing is, the greater the influence [16].

2.5. Logistic Regression

LR is a well-known algorithm that produces ordinal data rankings [0,1]. LR is a mathematical model that allows the estimation of the probability of having a particular class. As one of the

practical classification tools, LR is well-known for its vast range of applications, including machine learning, data mining, pattern recognition, medical science, and statistics [28]. The link between the dependent variable in binary data and independent factors in the form of intervals and categorical data is explained using LR [29]. Binary variables are variables that only have two categories, namely the category that states the event of success (Y = 1) and the category that expresses the event of failure (Y = 0) [30].

Consider a binary classification problem with d-dimensionality with feature vector $X = (X_1, X_2, X_3, \dots, X_p)$ and variable class $Y \in \Theta = \{\theta_1, \theta_2\}$. Let $p_1(x)$ show that the probability $Y = \theta_1$ is given $X = x$. Then, in the binary Logistic Regression model, it is assumed as follows:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_p x_p}} \quad (8)$$

Where $\beta \in R^d$, $\beta_0 \in \mathbb{R}$ is the parameter vector and β^T is the transpose of β. Transform $p(x)$ with the logit transformation g(x), so that:

$$g(x) = ln\left(\frac{p(x)}{1-p(x)}\right) \quad (9)$$

Generate logit form:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_1 x_p \quad (10)$$

2.6. Ensemble Learning

Ensemble Learning is a method that combines the predictions of several machine learning-based algorithms to make more accurate predictions. In other words, in ensemble learning, several learning models are trained to create a robust predictive model [31].

Given data with n samples and m features $D = \{(x_i, y_i)\}(|D| = n,\ x_i \in R^m, y_i \in R)$, the ensemble learning model uses an aggregation function G that combines K models, $\{f_1, f_2, \dots, f_k\}$ goes to predict a single output as follows:

$$\hat{y} = \varphi(x_i) = G(f_1, f_2, \dots, f_k) \quad (11)$$

Where $\hat{y} \in Z$ for classification problem. Based on this general framework, building a model ensemble involves selecting a methodology to train the participating models, selecting an appropriate process, and combining model outputs [32].

a. Voting

Voting is a machine learning model that practices on an ensemble of various models and predicts the output (class) based on the highest probability of the class being selected as output. Combining the findings of each classifier passed to the voting classifier and predicting the output class based on the highest majority of votes. The idea is to create separate custom models, find each model's accuracy, create a single model trained by this model, and predict the output based on the combined majority of votes for each output class. The selected class is selected based on the highest score in the total vector as follows [33]:

$$\hat{y} = argmax\big(\varphi(x_i)\big) \text{where } \varphi(x_i) = \sum_{k=1}^{K} f_k\,(x_i) \quad (12)$$

b. Averaging

In this method, the average prediction of all models is taken and used to make the final prediction. This approach only uses the average output of individual classifiers on different classifiers. The output that produces the maximum mean value is selected as the appropriate class. Generally, this method applies to the output classifier members that are numeric. Averaging can make predictions in regression problems or when calculating probabilities for classification problems [34].

c. Stacking

Stacking is an ensemble method that combines the outputs of heterogeneous base classifiers to improve prediction performance. An ensemble with a stacking technique consists of a base classifier and a meta-classifier. Each primary classifier is trained differently, using different learning algorithms to perform the target task. Meta classifiers are trained to combine the different strengths of heterogeneous base classifiers by determining which base classifier is more likely to be accurate for each class in carrying out the task. When an example is given, the individual base classifier classifies it independently. The output of the base classifier is then fed into the meta classifier to make the final prediction [35].

2.7. Model Evaluation

There are several evaluation models used in the QSAR classification, namely: accuracy, sensitivity,

specificity, and the Matthews Correlation Coefficient (MCC)[5].

The performance of the classification QSAR model is calculated based on the following equation:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (15)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (16)$$

Accuracy, sensitivity, specificity, and MCC are derived from the confusion matrix. Accuracy is the most popular measure for evaluating performance in classification because of its simplicity and importance. Accuracy is a representation of the ratio of the number of correctly classified samples and the total number of samples. Sensitivity is the proportion of correctly classified positive samples, while specificity represents the proportion of correctly classified negative samples. Finally, MCC is the correlation coefficient between observed and predicted in the classification [13].

3. RESEARCH METHOD

3.1. Research data

The research built the classification QSAR model in this thesis using data on the target of Alzheimer's disease drugs, namely acetylcholinesterase inhibitors (AChEIs). The data used in this study was obtained from the ChEMBL database, which can be accessed through the website www.ebi.ac.uk/chembl by selecting bioactivity for humans (CHEMBL220), which was accessed on 5 March 2021.

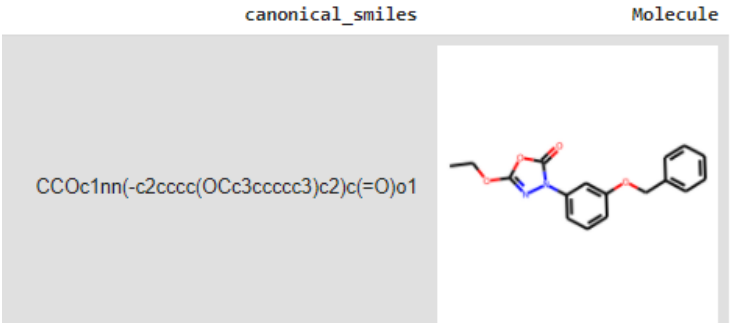
TABLE 1. Class Determination Category

No.	Activity	Category	Class
1	IC_{50}	(<1000 nM)	active
2	IC_{50}	(1000nM≤ IC_{50} ≤10000nM)	gray
3	IC_{50}	(< 10000nM)	inactive

At the stage of categorizing the data, there were 2215 molecules of active compounds, 1594 molecules of inactive compounds, and 1234 molecules of gray compounds. Molecules of compounds that fell into the gray class were discarded because they were not used to build a classification QSAR model [36]. Feature extraction steps were carried out with the help of the KNIME data analysis platform, and normalization and removal of salt and small fragments were performed on compound molecules using the "RDKit Salt Stripper" node. In the data preprocessing proceed, four datasets were generated in the form of fingerprints, with the specifications for each data shown in Table 2.

TABLE 2. Fingerprint Dataset Specification

No	Fingerprints	Row	Column
1	ECFP4	3809	1024
2	ECFP6	3809	1024
3	FCFP4	3809	1024
4	FCFP6	3809	1024



Fingerprints

1,1,1,0,0,0,0,0,0,1,1,1,1,0,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,…,0,0,0,0,0,0,0,0,0,0,0,0, 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0

FIGURE 3. Extract the compound and convert it into a molecular description.

3.2. Research Workflow

In this research, Google Colab (www.colab.research.google.com) was used to run Python 3.7.10 and KNIME 4.3.2 programming languages to remove salt and small fragments and then extracted

based on SMILES into a fingerprint. The steps of work carried out in this study are described as follows:

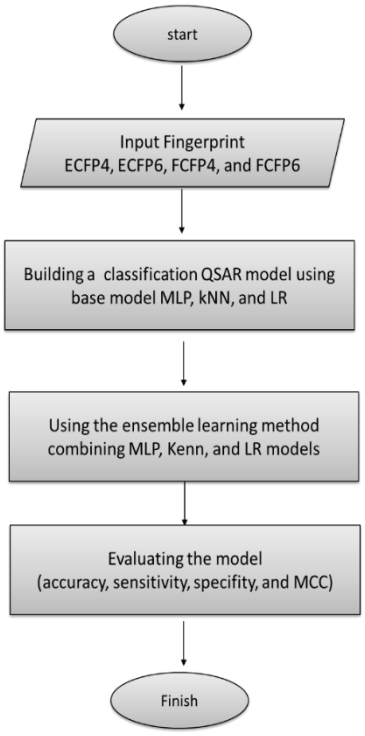


FIGURE 4. Research Workflow in This Study using Ensemble Learning

3.3. Ensemble Learning

The ensemble learning method is a method that combines several models to produce better model performance. In this study, the ensemble method combined the MLP, kNN, LR models, and the ensemble technique used voting, averaging, and stacking to determine the best parameters.

a. Voting

The ensemble learning method using the voting technique used a majority vote from the base models that would be the final prediction. The working steps for the voting technique can be seen in Figure 5.

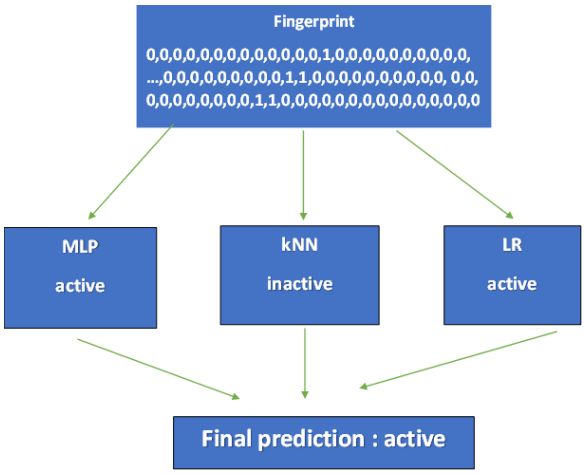


FIGURE 5. Voting Process

b. Averaging

The step for the ensemble learning method using the averaging technique was to use the average evaluation of the base models, which would be the final prediction. The working steps for the averaging technique can be seen in Figure 6.

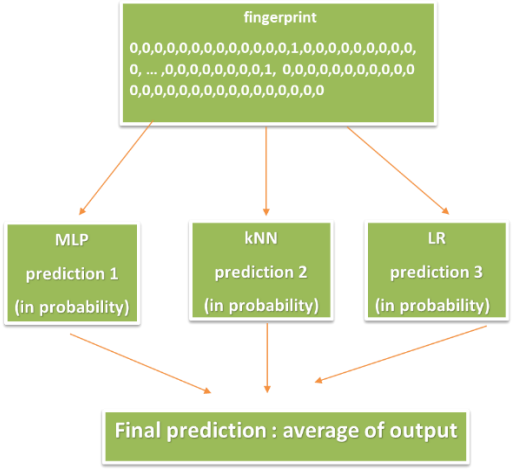


FIGURE 6. Averaging Process

c. Stacking

The ensemble learning method used the stacking technique to train first-level learners using the original training data set. Then, it generated a new data set to train second-level learners, where the output of the first-level learner was considered an input feature. In contrast, the original label was

still considered a new training data label. Next, a learning algorithm (second-level learner) was applied, which combined the base models' evaluation to become the final prediction.

Any learning method could be used to learn a second-level classifier. The Stacking framework is a versatile tool that may be applied to various circumstances. For example, to build the first-level features and translate the data into another feature space can use a variety of classifiers and learning algorithms.

In ensemble learning theory, base models are models that can be used as building blocks for generating more intricate models by merging many of them. Unfortunately, these fundamental models do not perform well on their own, either because they are biased (For instance, models with a low degree of freedom) or because they have too much volatility to be accurate (high degree of freedom models, for example). Ensemble approaches work to minimize the bias and variance of such weak learners by combining many of them into a stronger learner (or ensemble model) that performs better. This study used MLP, kNN, and LR as base learners and LR as second/ meta learner. The working steps for the stacking technique can be seen in Figure 7.

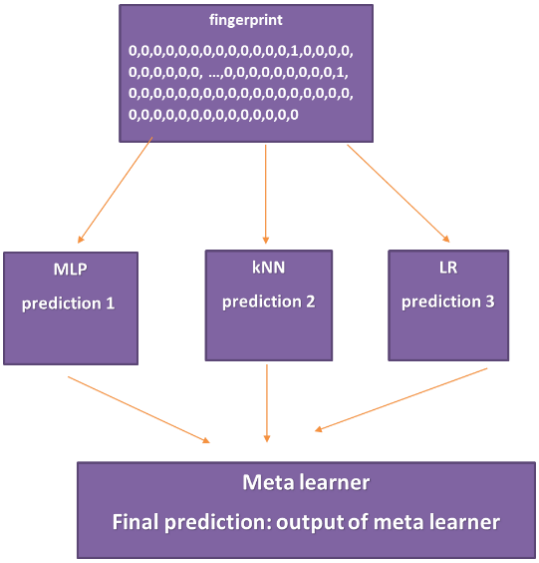


FIGURE 7. Stacking Process

4. Results and Discussion

This study built a classification QSAR model using the ensemble learning method that combined MLP, kNN, and LR models with voting, averaging and stacking techniques. The distribution of the 70-30, 80-20, and 90-10 datasets modified based on experimental observations of the proportion of data that have succeeded in increasing accuracy results in other research references in the classification QSAR model had better performance. Therefore, the dataset in this study was divided into 70% training data and 30% testing data, 80% training data and 20% testing data, and 90% training data and 10% testing data. The evaluation model used were accuracy, sensitivity, specificity, and MCC [5].

A. Comparison of QSAR Model Performance Using the Basic Model and Ensemble Learning Method

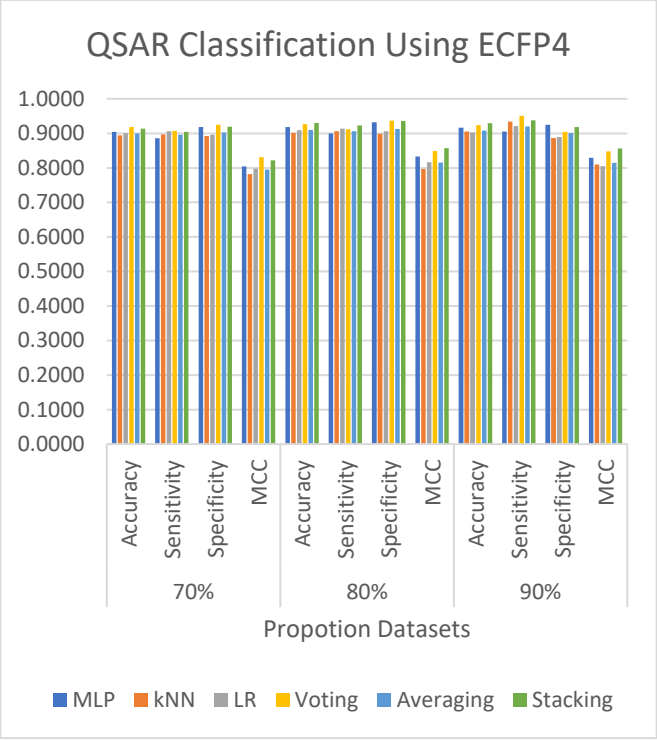


FIGURE 8. QSAR Model Performance Using ECFP4

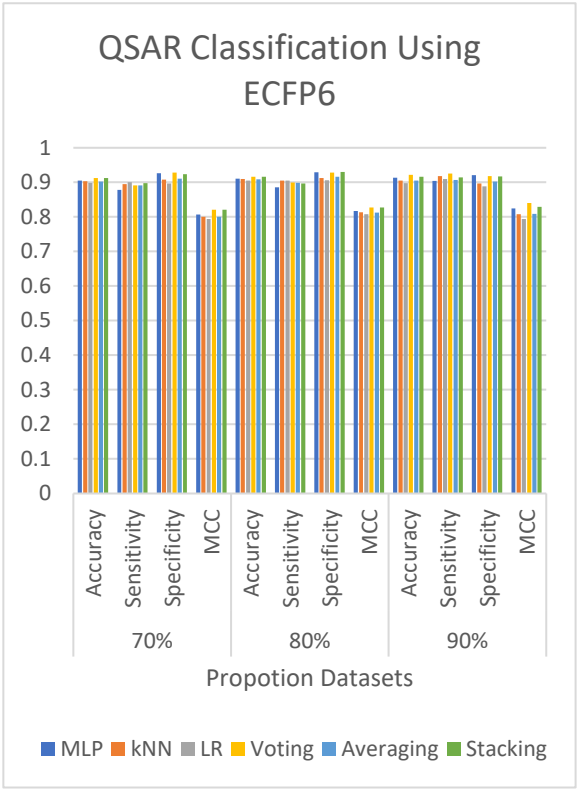


FIGURE 9. QSAR Model Performance Using ECFP6

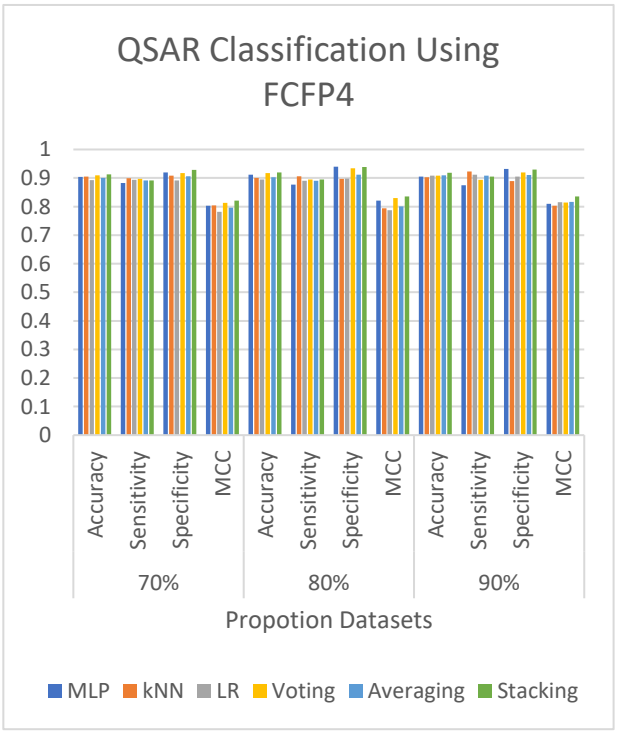


FIGURE 10. QSAR Model Performance Using FCFP4

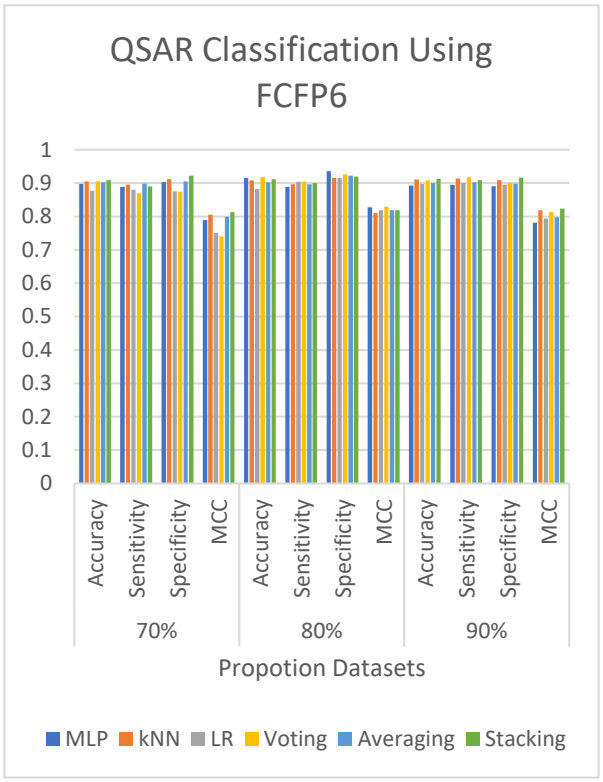


FIGURE 11. QSAR Model Performance Using FCFP6

The results indicate the ensemble learning method performed better for all fingerprint datasets than the base model. The ensemble learning method had better performance because it could overcome data imbalances and the Curse of dimensionality [37]. The model's performance is evaluated in the field using adequate data to ensure the research's legitimacy. There were more active inhibitors in this study than inactive data. Ensemble learning could also avoid overfitting and was more representative. The performance of the QSAR model using the ensemble learning method for the ECFP 4, FCFP4, and FCFP6 fingerprint datasets is better with the distribution of the 80-20 dataset compared to the distribution of other datasets. In the ECFP6 fingerprint dataset, the distribution of the 90-10 dataset was superior in several evaluations but also had several evaluations that had no better value than the distribution of the 80-20 dataset. Overall, the ensemble learning method for all fingerprint datasets with the distribution of the 80-20 dataset had better performance than the distribution of other datasets.

B. Comparison Of QSAR Model Performance Using Ensemble Learning Method

This section discusses the performance of the classification QSAR model with the ensemble learning method using the ECFP4, ECFP6, FCFP4, and FCFP6 fingerprint datasets. Each bar chart will show the performance of voting, averaging, and stacking techniques based on the classification evaluation.

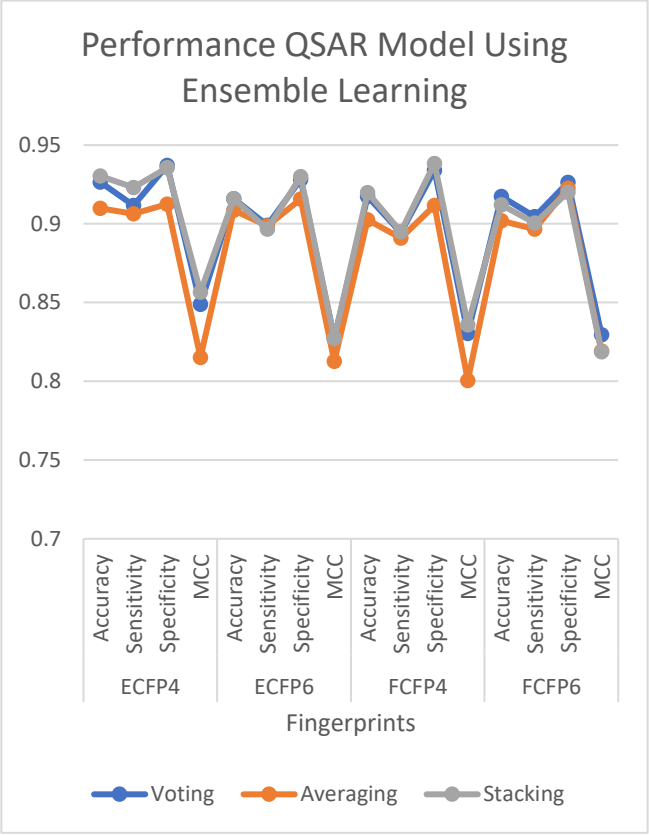


FIGURE 12. QSAR Model Performance Using Ensemble Learning Method

The ECFP4 dataset had the best performance for the classification QSAR model in this study. ECFP4 is a fingerprint method that is commonly used because it can represent compound molecules well [10]. This study showed that ECFP4 was a suitable fingerprint method for building a QSAR model with the ensemble learning method. This study was also in line with the results of research [6], where ECFP4 was the best hash fingerprint for DPP-IV. ECFP4 used for feature extraction of AChE inhibitors was a suitable fingerprint method used feature extraction because it

successfully represented the compound's molecular structure. It had the best performance for each ensemble learning technique used in this study.

The QSAR performance classification results using the ensemble learning method with the stacking technique had the best performance for the ECFP4, ECFP6, and FCFP4 datasets. In contrast, for the FCFP6 dataset, the voting technique has the best performance. By deciding which base classifier is more likely to be correct for each class in carrying out the task, the stacking strategy can combine the diverse strengths of the heterogeneous basic model. Stacking had the best performance among the classification QSAR model with the proposed ensemble learning method. In this study, the dataset with the best performance was ECFP4 with the ensemble learning method of stacking techniques. To provide better study results, the researcher should improve the current model by attempting to increase the number of proportions in the data to be investigated in future work.

5. CONCLUSION

New Acetylcholinesterase inhibitors (AChEI) should be discovered to be employed as low-risk treatments for Alzheimer's disease. This research aims to use machine learning and deep learning methods to create QSAR classification models.

There are often millions of possible therapeutic molecules under evaluation in the early stages of drug development and design. Therefore, it is critical to forecast drug candidate activity early by utilizing computational (in silico) methods to save time and resources. The activity of biological compounds can be predicted using QSAR. QSAR is an approach for building mathematical models that aid in understanding the relationship between a molecular compound's chemical structure and biological activity. Machine learning approaches that have been developed or developed in the realm of drug discovery can be used to implement QSAR.

Circular fingerprints were discovered to be particularly capable of describing molecular structure in this study, and ensemble learning may then be used to categorize the active or inactive structure of a molecule in the context of machine learning. These fingerprints improve the capacity of a

model created for these characteristics to generalize from training compounds to the novel, previously unknown structures. ECFP descriptors can represent many different features that can be interpreted as the presence of a particular substructure to simplify the analysis results. The performance of the classification QSAR model with Ensemble Learning, the best stacking technique, was owned by the circular fingerprints ECFP4 with the values of accuracy, sensitivity, specificity, and MCC respectively 93.04%, 92.58%, 93.36%, and 85.66%. ECFP4 dataset had the best performance for the classification QSAR model. This meant that ECFP4 could represent compound molecules well.

In this research, 80-20 data distribution has a better performance than the distribution of other datasets. This data sharing is commonly used in research because it produces better performance. For instance, QSAR classification using ECFP4 with ensemble learning performed with distribution 70 - 30 yielded accuracy, specificity, sensitivity and MCC respectively 91%, 90%, 92%, and 82% and for distribution 80 – 20 yielded 93%, 92%, 94%, and 86% whereas for distribution 90 – 10 yielded 92%, 93%, 91%, and 85%. The classification QSAR model that used the base model has lower performance results than using the ensemble learning method.

A model's sensitivity, specificity, and accuracy parameters should be balanced such that it can accurately discriminate between active and inactive substances. The QSAR classification model, which employs ensemble learning techniques, demonstrated a balanced sensitivity, specificity, and accuracy performance. Therefore, the authors would not only look at the balance of sensitivity, specificity, and accuracy values in this study. Still, they would also compare the best performance of the QSAR classification model using the MCC value, which was a more representative measuring metric than accuracy. The MCC value for the QSAR classification model using ensemble learning is 86%.

The selection of a good base model would affect the performance of the ensemble learning method because ensemble learning manages the strengths and weaknesses of each base model. The QSAR classification model using the ensemble learning method was proven to perform better than the base or single model. The stacking technique can combine the different strengths of a

heterogeneous base model by determining which base classifier is more likely to be accurate for each class in carrying out the task so that it has better performance than voting and averaging techniques. The highest performance of the classification QSAR model using the ensemble learning method was obtained by stacking techniques with an evaluation of the classification accuracy, sensitivity, specificity, and MCC, respectively 93.04%, 92.31%, 93.56%, and 85.66%. As a result, an ensemble learning classifier model can be utilized to model QSAR classification.

DATA AVAILABILITY

The dataset used for this study can access on website www.ebi.ac.uk/chembl.

ACKNOWLEDGEMENTS

To members of the Laboratory of Bioinformatics and Advanced Computing (BACL), Department of Mathematics and Data Science (DSC), Faculty of Mathematics and Natural Sciences, University of Indonesia, the authors are grateful for your support. This research was supported by PUTI Q2 2020 research grant from Directorate of Research and Development of Universitas Indonesia with contract number NKB-1645/UN2.RST/HKP.05.00/2020. Our special thanks to Enago (www.enago.com) for the English review of this paper.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] T.C. dos Santos, T.M. Gomes, B.A.S. Pinto, et al. Naturally occurring acetylcholinesterase inhibitors and their potential use for Alzheimer's disease therapy, Front. Pharmacol. 9 (2018), 1192. https://doi.org/10.3389/fphar.2018.01192.

[2] P. Scheltens, K. Blennow, M.M.B. Breteler, et al. Alzheimer's disease, The Lancet. 388 (2016), 505–517. https://doi.org/10.1016/s0140-6736(15)01124-1.

[3] M. Rohini, D. Surendran, Classification of neurodegenerative disease stages using ensemble machine learning classifiers, Procedia Computer Sci. 165 (2019), 66–73. https://doi.org/10.1016/j.procs.2020.01.071.

[4] A. Miličević, G. Šinko, Development of a simple QSAR model for reliable evaluation of acetylcholinesterase inhibitor potency, Eur. J. Pharm. Sci. 160 (2021), 105757. https://doi.org/10.1016/j.ejps.2021.105757.

[5] K. Roy, S. Kar, R.N. Das, A primer on QSAR/QSPR modeling, Springer International Publishing, Cham, 2015. https://doi.org/10.1007/978-3-319-17281-1.

[6] S. Syarofina, A. Bustamam, A. Yanuar, et al. The distance function approach on the MiniBatchKMeans algorithm for the DPP-4 inhibitors on the discovery of type 2 diabetes drugs, Procedia Computer Sci. 179 (2021), 127–134. https://doi.org/10.1016/j.procs.2020.12.017.

[7] Q. Li, X. Ding, H. Si, H. Gao, QSAR model based on SMILES of inhibitory rate of 2, 3-diarylpropenoic acids on AKR1C3, Chemometrics Intell. Lab. Syst. 139 (2014), 132–138. https://doi.org/10.1016/j.chemolab.2014.09.013.

[8] A. Worachartcheewan, P. Mandi, V. Prachayasittikul, et al. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors, Chemometrics Intell. Lab. Syst. 138 (2014), 120–126. https://doi.org/10.1016/j.chemolab.2014.07.017.

[9] S. Zhong, J. Hu, X. Fan, et al. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants, J. Hazardous Mater. 383 (2020), 121141. https://doi.org/10.1016/j.jhazmat.2019.121141.

[10] A. Cereto-Massagué, M.J. Ojeda, C. Valls, et al. Molecular fingerprint similarity search in virtual screening, Methods. 71 (2015), 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

[11] H. Feng, L. Zhang, S. Li, L. Liu, T. Yang, P. Yang, J. Zhao, I.T. Arkin, H. Liu, Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints, Toxicol. Lett. 340 (2021), 4–14. https://doi.org/10.1016/j.toxlet.2021.01.002.

[12] H. Hamzah, A. Bustamam, A. Yanuar, et al. Classification analysis using support vector machine, decision tree, and neural network with principal component analysis to determine molecular structure relationship from its biological activity on dipeptidyl peptidase IV inhibitors, AIP Conf. Proc. 2296 (2020), 020092. https://doi.org/10.1063/5.0030748.

[13] Mushliha, A. Bustamam, A. Yanuar, et al. Comparison accuracy of multi-layer perceptron and DNN in QSAR classification for acetylcholinesterase inhibitors, in: 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), IEEE, Bandung, Indonesia, 2021: pp. 1–6. https://doi.org/10.1109/AIMS52415.2021.9466040.

[14] P. Anki, A. Bustamam, R.A. Buyung, Comparative analysis of performance between multimodal implementation of chatbot based on news classification data using categories, Electronics. 10 (2021), 2696. https://doi.org/10.3390/electronics10212696.

[15] Y.S. Park, S. Lek, Artificial neural networks, in: Developments in Environmental Modelling, Elsevier, 2016: pp. 123–140. https://doi.org/10.1016/B978-0-444-63623-2.00007-4.

[16] G. Tuerhong, M. Wushouer, D. Zhang, An improved K nearest neighbor classifier for high-dimensional and mixture data, J. Phys.: Conf. Ser. 1813 (2021), 012026. https://doi.org/10.1088/1742-6596/1813/1/012026.

[17] S. Kang, k-Nearest neighbor learning with graph neural networks, Mathematics. 9 (2021), 830. https://doi.org/10.3390/math9080830.

[18] R. Wang, N. Xiu, S. Zhou, An extended Newton-type algorithm for ℓ_2-regularized sparse logistic regression and its efficiency for classifying large-scale datasets, J. Comput. Appl. Math. 397 (2021) 113656. https://doi.org/10.1016/j.cam.2021.113656.

[19] A. Field, Logistic regression logistic regression logistic regression, Discov. Stat. Using SPSS, 29 (2012), 731–735.

[20] N.E.H. Hammoudi, W. Sobhi, A. Attoui, et al. In silico drug discovery of Acetylcholinesterase and Butyrylcholinesterase enzymes inhibitors based on Quantitative Structure-Activity Relationship (QSAR) and drug-likeness evaluation, J. Mol. Struct. 1229 (2021), 129845. https://doi.org/10.1016/j.molstruc.2020.129845.

[21] J. Kumar, A. Gill, M. Shaikh, et al. Pyrimidine-triazolopyrimidine and pyrimidine-pyridine hybrids as potential acetylcholinesterase inhibitors for Alzheimer's disease, ChemistrySelect. 3 (2018), 736–747. https://doi.org/10.1002/slct.201702599.

[22] S. Das, M.A. Laskar, S.D. Sarker, et al. Prediction of anti-alzheimer's activity of flavonoids targeting acetylcholinesterase in silico, Phytochem. Anal. 28 (2017), 324–331. https://doi.org/10.1002/pca.2679.

[23] H. Yang, L. Sun, W. Li, et al. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts, Front. Chem. 6 (2018), 30. https://doi.org/10.3389/fchem.2018.00030.

[24] A. Bustamam, H. Hamzah, N.A. Husna, et al. Artificial intelligence paradigm for ligand-based virtual screening on the drug discovery of type 2 diabetes mellitus, J. Big Data. 8 (2021), 74. https://doi.org/10.1186/s40537-021-00465-3.

[25] M. Desai, M. Shah, An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN), Clinical eHealth. 4 (2021), 1–11. https://doi.org/10.1016/j.ceh.2020.11.002.

[26] S. Mirjalili, S.M. Mirjalili, A. Lewis, Let a biogeography-based optimizer train your Multi-Layer Perceptron, Inform. Sci. 269 (2014), 188–209. https://doi.org/10.1016/j.ins.2014.01.038.

[27] C. Nigam, A.K. Sharma, Experimental performance analysis of web recommendation model in web usage mining using KNN page ranking classification approach, Materials Today: Proceedings. In Press (2020). https://doi.org/10.1016/j.matpr.2020.09.364.

[28] R. Wang, N. Xiu, S. Zhou, An extended Newton-type algorithm for ℓ_2-regularized sparse logistic regression and its efficiency for classifying large-scale datasets, J. Comput. Appl. Math. 397 (2021), 113656. https://doi.org/10.1016/j.cam.2021.113656.

[29] P. Anki, A. Bustamam, R.A. Buyung, Looking for the link between the causes of the COVID-19 disease using the multi-model application, Commun. Math. Biol. Neurosci. 2021 (2021), 75. https://doi.org/10.28919/cmbn/6128.

[30] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, 3rd ed., Wiley, 2013. https://doi.org/10.1002/9781118548387.

[31] G.T. Reddy, S. Bhattacharya, S. Siva Ramakrishnan, et al. An ensemble based machine learning model for diabetic retinopathy classification, in: 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE), IEEE, Vellore, India, 2020: pp. 1–6. https://doi.org/10.1109/ic-ETITE47903.2020.235.

[32] S. Hakak, M. Alazab, S. Khan, et al. An ensemble machine learning approach through effective feature extraction to classify fake news, Future Gener. Comput. Syst. 117 (2021) 47–58. https://doi.org/10.1016/j.future.2020.11.022.

[33] O. Sagi, L. Rokach, Ensemble learning: A survey, WIREs Data Mining Knowl. Discov. 8 (2018), e1249. https://doi.org/10.1002/widm.1249.

[34] F. Huang, G. Xie, R. Xiao, Research on ensemble learning, in: 2009 International Conference on Artificial Intelligence and Computational Intelligence, IEEE, Shanghai, China, 2009: pp. 249–252. https://doi.org/10.1109/AICI.2009.235.

[35] H.M. Gomes, J.P. Barddal, F. Enembreck, et al. A survey on ensemble learning for data stream classification, ACM Comput. Surv. 50 (2018), 1–36. https://doi.org/10.1145/3054925.

[36] S. Simeon, N. Anuwongcharoen, W. Shoombuatong, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking, PeerJ. 4 (2016), e2322. https://doi.org/10.7717/peerj.2322.

[37] V. Kumar, S. Minz, Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification, Knowl. Inform. Syst. 49 (2016), 1–59. https://doi.org/10.1007/s10115-015-0875-y.