MALAYSIAN JOURNAL OF APPLIED SCIENCES 2023, VOL 8 (2): 28-36
E-ISSN:0127-9246 (ONLINE)
http://dx.doi.org/10.37231/myjas.2023.8.2.357
https://journal.unisza.edu.my/myjas

ORIGINAL ARTICLE

Prediction of Carboxylic Acid Toxicity Using Machine Learning Model

Zubainun Mohamed Zabidi[1*], Nurul Batrisyia Muhamad Suhaimy[1], Nur Diyana Nazihah Fuadi[1], Nur Hanisah Hamzi[1], Ahmad Nazib Alias[1]

Faculty of Applied Sciences, Universiti Teknologi MARA, Cawangan Perak, Kampus Tapah, 35400 Tapah Road Perak

*Corresponding author: zubainun384@uitm.edu.my

Received: 17/05/2023, Accepted: 30/10/2023, Available Online: 31/10/2023

Abstract

Carboxylic acids are organic compounds characterized by the presence of a carboxyl functional group capable of donating a proton and forming carboxylate ions in aqueous solutions. The carboxylic acid has widely been used in in manufacturing and medical applications. The rapid growth in carboxylic acid has established a need to predict its toxicity. The purpose of this paper to build predictive toxicity of carboxylic acid models by using five molecular descriptors (refractive index, The octanol/water partition coefficient (log P), acid dissociation constant (pKa), density, and dipole moment) through Machine Learning algorithms. The accuracy of the Machine Learning algorithm was determine by using three different types of models which are Decision Tree, Random Forest and k-Nearest Neighbour (k-NN). Among the machine learning algorithms used, we have determined that the decision tree is the best model for predicting the toxicity of carboxylic acid. This finding demonstrates that the decision tree model exhibits an acceptable level of performance in predicting toxicity within the field of toxicology.

Keywords: carboxylic acid; toxicity; machine learning

Introduction

Carboxylic acids are organic compounds characterized by the presence of a carboxyl functional group (-COOH) in their chemical structure. The carboxyl group consists of a carbonyl group (C=O) and a hydroxyl group (OH) bonded to the same carbon atom. Carboxylic acids can be either aliphatic or aromatic, depending on the nature of the carbon chain attached to the carboxyl group. They are typically acidic in nature, capable of donating a proton and forming carboxylate ions in aqueous solutions. Carboxylic acids are widely used for manufacturing plastic, cosmetics, and medicine. For instance, carboxylic acid is applied in manufacturing plasticisers and resins, raw materials of polyester, and the synthesis of nylon(Bhadra, Ahmed, Lee, & Jhung, 2022). Besides, in the medical field, ascorbic acid helps to maintain healthy skin, blood vessels, bone, and

cartilage, while benzoic acid acts as an antiseptic to treat urinary tract infections (Horgan & O'Sullivan, 2022). Furthermore, benzoic acid is also used in facial cleansers(Bayo et al., 2017).

Carboxylic acids can present various hazards depending on their specific chemical properties, concentration, and exposure conditions (Jos et al., 2009). Hazards associated with carboxylic acids include corrosive effects on skin, eyes, and mucous membranes due to their acidic nature(Jiang et al., 2019). Carboxylic acids can exhibit varying degrees of drug toxicity depending on factors such as their chemical structure, concentration, and mode of interaction with biological systems (Mitra, 2022). While some carboxylic acids are well-tolerated and even utilized therapeutically, others may pose toxicity concerns. The toxicity of carboxylic acids can arise from their ability to disrupt cellular processes, interfere with enzyme activity, or cause adverse effects on organs and tissues. The median lethal dose (LD50), used in toxicology, measures the amount of toxin, pathogen, or radiation needed to kill 50% of the test population within the text window. Determination of toxicity might involve animal testing such as in vivo and in vitro method which required high costs. Animal testing will require many animals to be tested which demand the ethics issue in the research. Furthermore, In vivo method takes longer to arrive at the result and requires a high equipment cost and a large amount of material needed (Price, Blagg, Jones, Greene, & Wager, 2009). Meanwhile, in vitro is carried out outside a living organism, whether in a test tube or culture dish (Blomme & Will, 2016). The usage of animals and chemicals in this experiment involving the controversy with the environment and physiologically limited.

In silico method is an alternative method to study the toxicology of chemical compound. in silico methods are based on the computational techniques to study and predict various properties and behaviours of chemicals toxicity (Borrero, Guette, Lopez, Pineda, & Castro, 2020). In silico also known as computational toxicology which employs computational resources such as methods, algorithms, software, or data-based to perform various tasks related to the assessment of toxicity (Limbu, Zakka, & Dakshanamurthy, 2022). These tasks include organizing, analyzing, modelling, simulating, visualizing, and predicting the toxicity of chemicals. Machine learning has emerged as a powerful tool for enhancing in silico methods, enabling the prediction and assessment of various properties and behaviours of chemical compounds. By leveraging computational algorithms and models, machine learning can analyze large datasets and identify patterns that correlate with toxicity. Implementing machine learning in in silico methods allows for faster and more accurate predictions of compound toxicity, aiding in the screening and prioritization of chemicals for further experimental testing (Vo, Van Vleet, Gupta, Liguori, & Rao, 2020). This integration of machine learning and in silico techniques holds promise for improving the efficiency and effectiveness of toxicology assessments in diverse scientific domains.

Utilizing the in silico method, machine learning can play a crucial role in diminishing the expenses associated with costly and invasive animal testing during clinical trials while determining toxicity. In this study focuses on predicting toxicity in carboxylic acid, employing molecular descriptors such as refractive index, The octanol/water partition coefficient (log P), acid dissociation constant (pKa), density, and dipole moment. By incorporating machine learning techniques, we aim to provide a more cost-effective alternative to traditional animal testing, reducing both financial burdens and ethical concerns in toxicity assessment.

Methodology

Data set

The chemical data of Carboxylic acids such as refractive index, pKa, density and dipole moment were collected using the Reaxys database. Reaxys offers powerful search and retrieval capabilities, allowing users to explore the vast chemical and scientific knowledge from different sources brought together in one central location (Ahmad Nazib Alias, Zabidi, Zakaria, Mahmud, &

Ali, 2021). The toxicology data was collected from PubChem data based. PubChem is a freely accessible database maintained by the National Center for Biotechnology Information (NCBI), which is a part of the United States National Library of Medicine (NLM)(Kim et al., 2018). In this study, the toxicology values were extracted from a single reference to ensure a compatible relationship of the data. The data utilized in this study are provided in the appendix section. The log P also was retrieved from PubChem database.

Data Pre-processing

The molecular descriptors such as refractive index, log P, pKa, density, and dipole moment was rescaling the data to a specific range between 0 and 1. This process is known as the normalization process. The purpose of normalization is to eliminate the impact of different scales or magnitudes of features, ensuring that no single feature dominates the analysis or modeling process based on its magnitude alone.

It can help improve the performance and convergence of machine learning algorithms, particularly those that are sensitive to the scale of the input data (e.g., gradient-based optimization algorithms). The input data were normalized using equation (1).

$$I_i = \frac{I_x - I_{min}}{I_{max} - I_{min}} \quad (1)$$

where I_x unnormalized input data, I_{max} was the maximum value of the sample and I_{min} was the minimum value of value of the sample.

In this work, we also using bootstrap resampling technique. Bootstrap is a powerful technique used in data preprocessing to estimate variability and assess uncertainty in a dataset. By randomly sampling observations from the original dataset with replacement, multiple bootstrap samples are created, capturing the inherent variability within the data. These bootstrap samples allow for the analysis of statistical measures, construction of confidence intervals, and validation of models by accounting for the uncertainty in the data. Bootstrap resampling plays a valuable role in understanding the robustness of results, identifying outliers, and evaluating the stability of statistical estimates in data preprocessing.

Machine learning Model.

In machine learning method, the data was split into training and test set as shown in figure 1. Splitting the dataset into training and testing sets is a crucial step in machine learning to evaluate the performance and generalization ability of the trained model. The main dataset will be split into two different data partitions, namely 80% of the training sets and 20% of the test sets.

We utilize RapidMiner Studio version 9.10 for conducting machine learning modeling and analysis. RapidMiner is a powerful and intuitive data science platform that enables users to perform a wide range of tasks in data analytics, machine learning, and predictive modeling. RapidMiner's intuitive Graphical User Interface (GUI) streamlines the data loading process. While, the predictive modelling is performed using the building operator. In this paper we use three predictive modeling that is k-nearest neighbor (k-NN) regression, decision tree, and random forest.

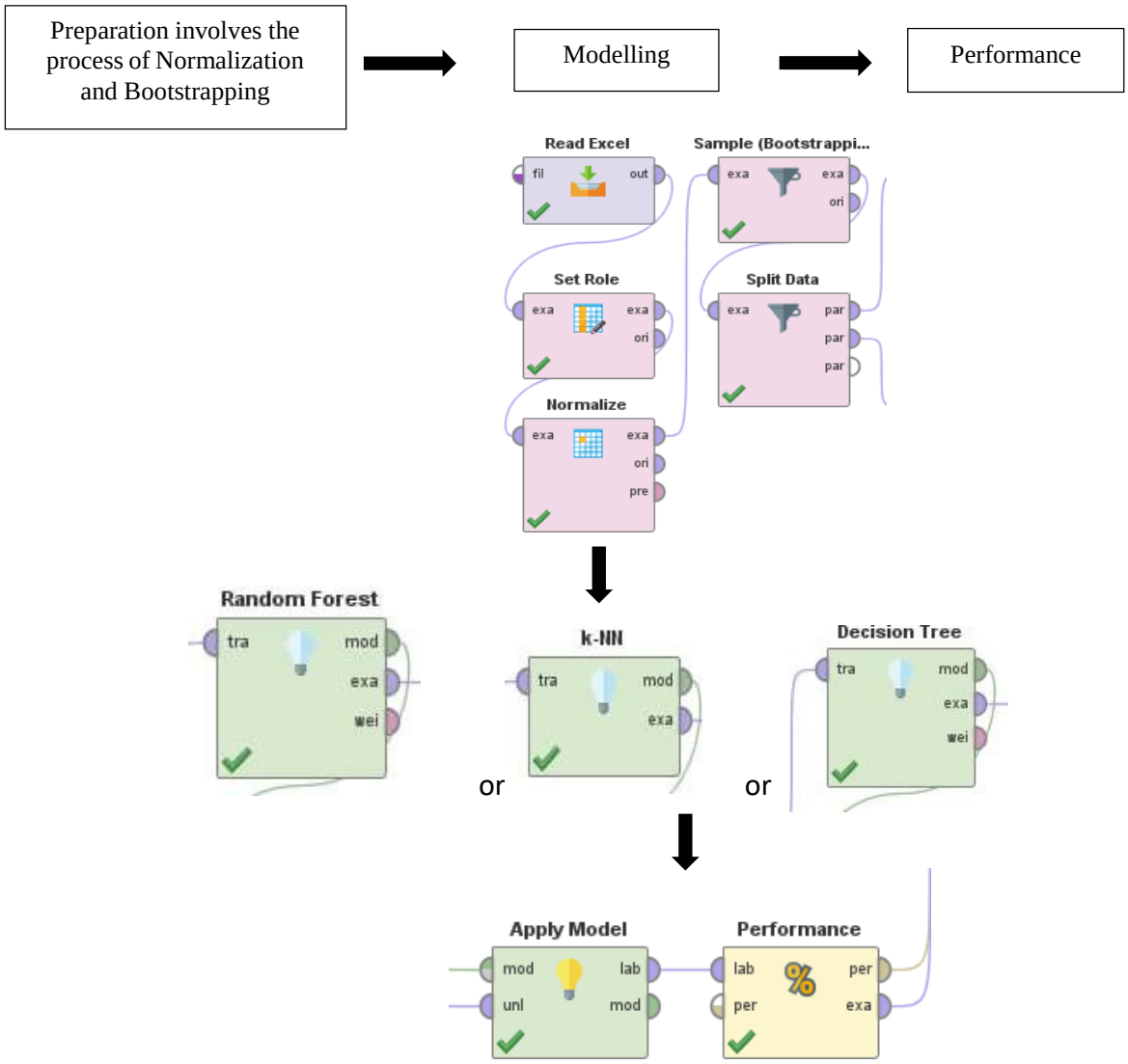

Figure 1. The machine learning model for predicting toxicology of the Carboxylic acids

k-NN model. The k-Nearest Neighbors (k-NN) algorithm is a popular and intuitive supervised machine learning method used for classification and regression tasks. It is a non-parametric algorithm that makes predictions based on the similarity between data points. In k-NN, the "k" represents the number of nearest neighbors used to make predictions. The algorithm classifies or predicts the target variable of a new data point by considering the class or value of its nearest neighbors in the feature space. k-NN is known for its simplicity, interpretability, and ability to capture complex decision boundaries. It is particularly useful when the data lacks a discernible underlying distribution or when interpretability is a priority. k-NN does not involve explicit training as it memorizes the entire training dataset, making prediction efficiency dependent on the size of the dataset.

Decision Tree model. The Decision Tree algorithm is a popular supervised machine learning technique used for classification and regression tasks. Figure 2 shows the constructs in hierarchical structure of decision tree. The root node is the starting point of the tree, representing the initial decision based on a specific feature. Each internal node represents a decision based on a specific feature or attribute. The leaves of the tree represent the final predictions or outcomes.

Decision Trees utilize a series of if-else conditions at each node to guide the traversal through the branches, ultimately reaching a leaf node. Each leaf node holds a specific prediction or class label. The root node captures the most influential feature that best splits the data. The branches or edges between nodes indicate the flow of decisions based on the feature values. The depth of the tree corresponds to the number of decision levels from the root to the leaves.

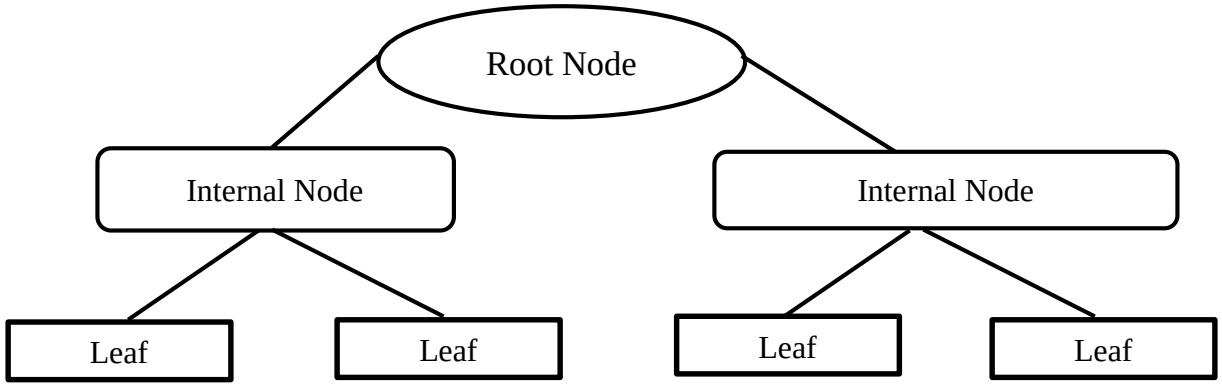

Figure 2. Basic Structure of decision tree.

Random Forest. Random Forest is an ensemble learning method in machine learning that utilizes multiple Decision Trees to make predictions. Each Decision Tree in the Random Forest is trained on a different subset of the data and with a random selection of features. The trees in a Random Forest share a common root but differ in the splits and outcomes at the leaf level. The combination of these individual trees creates a robust and accurate model. Each tree within the Random Forest has its own root, branches, nodes, leaves, and decision-making process. The root node of each tree represents the starting point for making decisions based on specific features. The branches and nodes guide the traversal through the tree, while the leaves represent the final predictions or outcomes. The root, leaf, and branching concepts from Decision Trees are preserved in Random Forest, as each tree in the ensemble follows the same hierarchical structure. The difference lies in the diversity of trees, each capturing a different subset of features and instances, leading to a more comprehensive and accurate prediction.

Machine Learning Performance

The root-mean-squared error (RMSE) is a commonly used metric for evaluating the performance of machine learning models. It measures the average magnitude of the differences between predicted and actual values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}\right)^2} \tag{2}$$

where N is the number of data, y_i is the actual values and $\hat{y}$ is prediction (calculated) value.

The Absolute Error is a metric used to measure how close the predictions of a machine learning model are to the actual values. It calculates the average difference between the predicted values and the experimental values.

$$\text{Absolute error} = \frac{1}{N}\sum_{i}^{N} | \ y_i - \hat{y} \ | \tag{3}$$

The value of *r* was computed using equation (4)

$$r = \frac{\Sigma\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\Sigma\left(x-\bar{x}\right)^2}\sqrt{\Sigma\left(y-\bar{y}\right)^2}} \tag{4}$$

where $\bar{x}$ and $\bar{y}$ is the mean value of independent descriptor, *x* and predicted, *y* respectively.

The R-squared (R^2) value, also known as the coefficient of determination, is a statistical metric used to assess the goodness-of-fit of a linear relation between predicted and experimental value.

Results and Discussion

In order to obtain the most accurate predictive models, a variety of approaches such as k-NNs, decision trees, and random forests are utilized in the machine learning approach. The machine learning approach involves the utilization of programming techniques to train computers in optimizing their performance and criteria. This optimization process is crucial as it contributes significantly to achieving accuracy in predictive models and ensuring the fitness of attribute descriptors namely refractive index, log P, pKa, density, and dipole moment. By employing data-driven predictive models, we gain insights into the toxicology model based on molecular descriptors. These models allow us to explore the relationships between the characteristics of chemical compounds and their toxicological properties.

Table 1. List of results obtained in the different models

MODEL	PERFORMANCE	DESCRIPTOR Refractive index and Dipole moment Test Value
Decision Tree	RMSE	0.222
	Absolute Error	0.189
	r	0.923
	R^2	0.853
Random Forest	RMSE	0.443
	Absolute Error	0.265
	r	0.656
	R^2	0.430
k-NN (k-Nearest Neigbour)	RMSE	0.535
	Absolute Error	0.375
	r	0.392
	R^2	0.154

In this work, three types of models have been built which are Decision Tree, Random Forest, and k-Nearest Neighbour (k-NN). All models have different test set accuracy as shown in table 1. The performance of the test data directly impacts the evaluation and assessment of a machine learning model. The test data serves as an independent sample that measures the model's ability to generalize and make accurate predictions on data model. According to the results above, we obtained the test value of RMSE, absolute error, r, and R^2. We find that, the best performance is decision tree model, the value RMSE, absolute error, r and R^2 is 0.222, 0.189, 0.923 and, 0.852 respectively. The high performance of test data, indicating good accuracy and generalization, it provides confidence in the model's effectiveness and reliability. This is show that the decision tree model has acceptable model for LD50. This finding demonstrates that the decision tree model exhibits an acceptable level of performance in predicting toxicity within the field of toxicology. The decision tree model successfully captures and represents the underlying relationships and patterns between the descriptors and the toxicological outcomes of carboxylic acid. It demonstrates a reasonable level of accuracy and reliability in its predictions, showcasing its efficacy as a tool for toxicology modeling. Meanwhile for Random Forest, the test value is 0.443,0.265,0.656 and 0.430 respectively. The test performance for k-NN is 0.535, 0.375,0.392 and 0.154 respectively.

To strengthen our statistical modeling, we also calculate the predictive criteria R^2 value of the log (LD50) of the experimental and calculated values for each molecular descriptor. Figure 3 shows the experimental and calculated log (LD50) for the decision tree algorithm. The graph for experimental vs calculated value allows us to visualize the comparison of the calculated value using decision tree model with the corresponding experimental values. In decision tree we find that, the calculated values align closely with the experimental values, which indicates that the model has good accuracy in predicting the toxicological from five descriptors refractive index, log P, pKa, density, and dipole moment. The experimental and calculated log (LD50) for the random forest and k-NN is given in the supplementary.

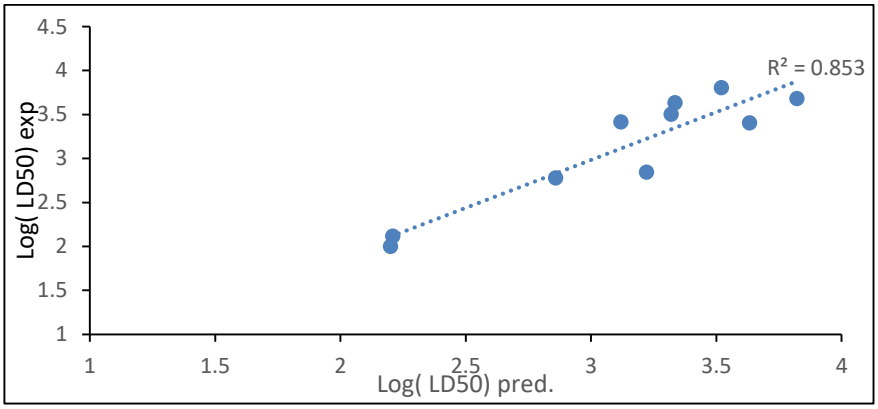

Figure 3. the experimental and calculated log (LD50) for the decision tree algorithm

The pKa value of a compound determines its degree of ionization at different pH levels. The pKa of a compound can affect its reactivity or ability to undergo chemical transformations. Reactive compounds with extreme pKa values may exhibit specific toxicological properties due to their reactivity with biological molecules or cellular components(Mansouri et al., 2019). pKa affects absorption, distribution, metabolism and excretion in toxicity properties(Zhang, 2006). The octanol/water partition coefficient (log P) is a measure of a compound's hydrophobicity and its tendency to partition between octanol (a nonpolar solvent) and water (a polar solvent). Log P provides information about the compound's lipophilicity, which is the tendency of a compound to dissolve in lipid-based environments. Lipophilicity is an essential property for understanding the

compound's absorption, distribution, and interaction with biological systems which importance parameter in toxicity analysis(Gaillard, Carrupt, Testa, & Boudon, 1994).

Molecular density commonly has been used as molecular descriptor because it provides basic information about the size and complexity of a chemical compound which aiding in the understanding of structure-activity relationships. The molecular dipole moment also an importance molecular descriptor in structure-activity relationships due to these properties can give important aspects of compound polarity, solubility, permeability, intermolecular interactions, stability, and reactivity(Ahmad NAZİB Alias & Zabidi, 2022). It aids in understanding the structure-activity relationships and provides valuable insights into the compound's behavior in biological systems. The refractive index in molecule related with to the polarizability of a molecule, which reflects its ability to undergo polarization in the presence of an electric field(Ahmad NAZİB Alias & Zabidi, 2022). Polarizability can affect intermolecular interactions, solubility, and other properties relevant in in silico study.

Conclusion

In this paper we have investigating three machine learning model namely decision tree, k-NN and random forest for in silico study. Five molecular descriptors have been used for machine learning toxicology modelling viz. dipole moment, refractive index, density, log P and pKA. In handling small database, we used bootstrap to increase stability and robustness of the data. We find that in our work the Decision Tree is the best machine learning algorithm that can be used to predict the toxicity of carboxylic acid. It followed by random forest and k-NN model. The high prediction rate was caused by these three properties. The selection of molecular descriptors and the choice of a suitable learning modeling algorithm are crucial factors in conducting an effective in silico study. These choices directly impact the accuracy of machine learning model which has the ability to gain meaningful insights from the analysis.

Acknowledgments

The authors would like to thank Faculty of Applied Sciences, Universiti Teknologi MARA, Cawangan Perak, Kampus Tapah for fully support in this works.

References

Alias, A. N., & Zabidi, Z. M. (2022). QSAR Studies on Nitrobenzene Derivatives using Hyperpolarizability and Conductor like Screening model as Molecular Descriptors. *Journal of the Turkish Chemical Society Section A: Chemistry, 9*(3), 953-968.

Alias, A. N., Zabidi, Z. M., Zakaria, N. A., Mahmud, Z. S., & Ali, R. (2021). Biological activity relationship of cyclic and noncyclic alkanes using quantum molecular descriptors. *Open Journal of Applied Sciences, 11*(8), 966-984.

Bayo, J., Martínez, A., Guillén, M., Olmos, S., Roca, M.-J., & Alcolea, A. (2017). Microbeads in Commercial Facial Cleansers: Threatening the Environment. *CLEAN – Soil, Air, Water, 45*(7), 1600683.

Bhadra, B. N., Ahmed, I., Lee, H. J., & Jhung, S. H. (2022). Metal-organic frameworks bearing free carboxylic acids: Preparation, modification, and applications. *Coordination Chemistry Reviews, 450*, 214237.

Blomme, E. A. G., & Will, Y. (2016). Toxicology Strategies for Drug Discovery: Present and Future. *Chemical Research in Toxicology, 29*(4), 473-504.

Borrero, L. A., Guette, L. S., Lopez, E., Pineda, O. B., & Castro, E. B. (2020). Predicting Toxicity Properties through Machine Learning. *Procedia Computer Science, 170*, 1011-1016.

Gaillard, P., Carrupt, P.-A., Testa, B., & Boudon, A. (1994). Molecular Lipophilicity Potential, a tool in 3D QSAR: Method and applications. *Journal of Computer-Aided Molecular Design, 8*(2), 83-96.

Horgan, C., & O'Sullivan, T. P. (2022). Recent developments in the practical application of novel carboxylic acid bioisosteres. *Current Medicinal Chemistry, 29*(13), 2203-2234.

Jiang, B., Gong, Y., Gao, J., Sun, T., Liu, Y., Oturan, N., & Oturan, M. A. (2019). The reduction of Cr(VI) to Cr(III) mediated by environmentally relevant carboxylic acids: State-of-the-art and perspectives. *Journal of Hazardous Materials, 365*, 205-226.

Jos, A., Pichardo, S., Puerto, M., Sánchez, E., Grilo, A., & Cameán, A. M. (2009). Cytotoxicity of carboxylic acid functionalized single wall carbon nanotubes on the human intestinal cell line Caco-2. *Toxicology in Vitro, 23*(8), 1491-1496.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., . . . Bolton, E. E. (2018). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research, 47*(D1), D1102-D1109.

Limbu, S., Zakka, C., & Dakshanamurthy, S. (2022). Predicting Dose-Range Chemical Toxicity using Novel Hybrid Deep Machine-Learning Method. *Toxics, 10*(11), 706.

Mansouri, K., Cariello, N. F., Korotcov, A., Tkachenko, V., Grulke, C. M., Sprankle, C. S., . . . Williams, A. J. (2019). Open-source QSAR models for pKa prediction using multiple machine learning approaches. *Journal of Cheminformatics, 11*(1), 60-71.

Mitra, K. (2022). Acyl Glucuronide and Coenzyme A Thioester Metabolites of Carboxylic Acid-Containing Drug Molecules: Layering Chemistry with Reactive Metabolism and Toxicology. *Chemical Research in Toxicology, 35*(10), 1777-1788.

Price, D. A., Blagg, J., Jones, L., Greene, N., & Wager, T. (2009). Physicochemical drug properties associated with in vivo toxicological outcomes: a review. *Expert Opinion on Drug Metabolism & Toxicology, 5*(8), 921-931.

Vo, A. H., Van Vleet, T. R., Gupta, R. R., Liguori, M. J., & Rao, M. S. (2020). An Overview of Machine Learning and Big Data for Drug Toxicity Evaluation. *Chemical Research in Toxicology, 33*(1), 20-37.

Zhang, H. (2006). A QSAR Study of the Brain/Blood Partition Coefficients on the Basis of pKa Values. *QSAR & Combinatorial Science, 25*(1), 15-24.

How to cite this paper:
Zabidi, Z. M., Muhamad Suhaimy, N. B., Nazihah Fuadi, N. D., Hamzi, N. H., Alias, A. N. (2023). Prediction of Carboxylic Acid Toxicity Using Machine Learning Model. *Malaysian Journal of Applied Sciences, 8(*2), 28-36.