Materials Research Proceedings 29 (2023) 234-247 https://doi.org/10.21741/9781644902516-27

Application of machine learning models to predict ecotoxicity of ionic liquids (Vibrio fischeri) using VolSurf principal properties

GRACE Amabel Tabaaza[1,a], BENNET Nii Tackie-Otoo[2,b], DZULKARNAIN B Zaini[1,c] and BHAJAN Lal[1, d *]

[1]Chemical Engineering Department, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia

[2]Petroleum Engineering Department, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia

[a]grace_20000207@utp.edu.my, [b]bennet_17006974@utp.edu.my, [c]dzulkarnain.zaini@utp.edu.my, [d]bhajan.lal@utp.edu.my

Keywords: Ionic liquids, Vibrio Fischeri, Toxicity, Machine learning, QSPR/QSAR and Principal Properties

Abstract. Owing to the rapid growth in IL synthesis due to feasible cation–anion combinations, knowledge of their toxicity is pertinent for their successful application. Toxicity information measurement of various ILs on a broad spectrum of conditions through experimental techniques is way demanding on time, resources, and is at times impractical. Various research works have been performed in Quantitative Structure Activity/Property Relationship (QSAR/QSPR) for IL toxicity prediction. In this study, ML models have been trained and tested on Vibrio fischeri toxicity data set using in silico principal properties (PPs) as descriptors. Deploying this properties aid in considering both the effect of cations and anions on *Vibrio fischeri* toxicity prediction. Among the models trained, the Random Forest model proved to be the most precise nevertheless, decision tree model was the most accurate and consistent. Considering the importance of the descriptors to *Vibrio fischeri* toxicity selection techniques and model optimization.

Introduction

Ionic liquids have unique characteristics viz low melting point, low vapor pressure and cation and anion composition, which allow them to serve as solvents, catalysts [1], [2], electrolytes and separating agents for numerous industrial solutions such as gas hydrate inhibition in Oil and gas pipeline[3]–[6] and promotion in carbon sequestration [7]–[9]. Ionic Liquids have low vapor pressure and may be regarded as "green" solvents. They are, therefore, not expected to have high toxicity in comparison with conventional volatile solvents in the environment. However, due to their high solubility and stability in water, they are likely to persist in wastewater. Therefore, it is essential to determine the level of risk to the aquatic life to successfully use these ILs.

Toxicity assessments of ILs are usually carried out on a marine luminescent bacterium, *Vibrio fischeri*. This is a photobacterium fischeri which emits light through normal metabolic process. When exposed to contaminants or pollutants, their metabolic process is affected which reflects as reduction in the amount of light emitted. This is a measure of ecotoxicity expressed as EC50 [10]–[16]. The international standard ecotoxicological bioassay (DINEN ISO11348) [17] recommends this bioassay with the bioluminescent bacterium *Vibrio fischeri* and is often considered as the reagent mostly used for toxicity determination [18]. Studies on environmental toxicity have adopted this assessment for testing chemicals [19]. This is attributed to the fact that bioassays with *Vibrio fischeri* are simple and fast in comparison to other procedures. Furthermore, this toxicity evaluation is applicable for many ILs because it shows optimum hydrophilic/lipophilic balance which is attributed to the fact that *Vibrio fischeri* is a gram-negative bacterium [16], [20]–[23].

Experimental toxicity measurement is the most effective and direct way of finding ILs with desired low toxicity. In addition, synthesis of IL structures that are novel is rapidly on the rise as a result of many feasible cation–anion combinations. Due to this rapid growth, toxicity measurement of various ILs under extensive conditions through experimental techniques requires lots of time and resource and are impractical. Therefore, various studies have been performed in Quantitative Structure Activity/Property Relationship (QSAR/QSPR) for IL toxicity prediction [24]. The models are developed using data on their properties acquired through experiments and this data can produce predictions on the toxicities of new ILs that are satisfactory. These studies use the "univariate" approach. In this approach, the cation effect is considered when the anion is fixed, and the anion effect is considered when the cation is fixed; assuming there is no toxicity variation with changing cationic or anionic IL counterparts. The other assumption is that there is no effect from the interaction between the cation and anion. Some studies also used electrostatic and topological structural descriptors or heuristic descriptors [25], [26].

To improve upon this approach and overcome the difficulty in the acquisition of experimental descriptors, Paternò et al [27] used GRID approach in VolSurf+ to derived the in silico cation and anion physicochemical descriptors. The VolSurf+ descriptors calculate interaction energy moments and capacity factors, assess hydrophilic and hydrophobic regions, molecular size and shape, compute amphiphilic moments, hydrophobic-lipophilic balance, as well as partition coefficient in different solvents, diffusion in water solvent, molecular flexibility in different solvents and pH dependent water solubility. A QSPR model was first used to validate the VolSurf+ descriptors using aquatic toxicity scores as responses. A good correlation was achieved yet there was difficulty in handling large number of descriptors especially for big data set [28]. Therefore, they developed nine principal properties (PPs) as physicochemical descriptors for 38 anions (4 PPs-) and 218 cations (5 PPs+) [27]. These PPs were used to come up with QSPR models for assessing and predicting IPC-81 rat cell line cytotoxicity, acetylcholinesterase inhibition [27] and *Vibrio fischeri* toxicity [16].

In this current study, the QSPR approach is extended by deploying other machine learning (ML) to predict *Vibrio fischeri* toxicity using the PPs developed by Paternò et al. [27] on the same IL data set used. In this study the cationic PPs are denoted as PP(n)C and the anionic ones are denoted as PP(n)A where the "n" is the numbering of the PPs. Five different ML models are trained and tested in this study. Their performances on the test data are compared to choose the best model.

Models used for supervised machine learning

Numerous industries have adapted supervised machine learning methods to solve complicated problems. In this current study QSPR studies has been extended by training and testing five ML models on *Vibrio fischeri* toxicity data using in silico descriptors developed by Paternò et al. [16], [27]. The five supervised ML models deployed in this study are briefly described below. Detailed description could be found in [29].

Decision tree models are used in machine learning to unite a series of the basic test efficiently where a numeric feature is compared to a threshold value in each test. Decision trees are an easy to use and well-known approach for statistical learning; such trees aim to identify the splitting criteria which describes the relationship between a set of input combinations, and regions of the output space. It is used to solve both regression and classification problems [30]–[32].

Random Forest (RF) model is usually used on datasets that are randomly sampled and functions by training decision trees on the datasets to have their predictions averaged. Features are generated from the predictions made by trees, and prediction made by one tree automatically becomes a feature for the prediction of the final model. It achieves good model results because it is easy to interpret and is robust against overfitting. RF is biased against features which are highly cardinal.

Extra tree regression: Geurts et al. [33] proposed that the random forest regression model has extra trees as extensions. These extra trees belong to the class of decision tree-based ensemble

learning methods. Multiple decision trees are used to perform classification and regression tasks in decision tree-based ensemble methods [34].

Gradient boosting (GB) regressor combines simple parameterized functions with "poor" performance (high prediction error) using an iterative algorithm in order to come up with a prediction rule that is highly accurate. Contrary to other methods of statistical learning which usually provide comparable accuracy (such as support vector machines and neural networks), GB produces results that are interpretable, and this is done without requiring a lot of data preprocessing and parameter tuning. GB is a technique used in machine learning for both regression and classification problems and produces an ensemble of weak prediction models simply called a prediction model [35].

Extreme Gradient Boosting (XGBoost) regression is carried out by means of the stochastic gradient boosting algorithm. It is a portable, flexible and efficient model for machine learning [36]. The stochastic gradient boosting algorithm has a parallel boosting technique that allows efficient, accurate and fast machine learning. As reported by Chen and Guestrin [36], XGBoost has one desired attribute of maximization of loss function. This helps prevent overfitting when the final weights are smoothened out and an extra regularization term is added to it. The successful performance of this model is due to its scalability in all setups. This model prevents over-fitting and shows great robustness against multicollinearity. It penalizes the irrelevant input variables coefficients, drawing these closer to zero and eliminating these zeroed inputs to minimize standard errors. This process allows for optimization of the algorithms.

Criteria for model evaluation

Models were evaluated using error matrices which indicate the deviation from true probabilities as well as the time taken to evaluate the models. These are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Akaike Information Criteria.

Mean Absolute Error (MAE): This indicates the deviation from true probability and is the mean of the absolute value of the errors. Mathematically expressed as,

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \quad (1)$$

Root mean Squares Error (RMSE): This is interpretable as the standard deviation of the prediction errors, a popular performance evaluation metric for models [37].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}, \quad (2)$$

Akaike Information criteria (AIC): To determine which of the many models is the best for dataset at hand, AIC is the single number score to use. The criteria estimates the model quality for each model relative to other models [38]. It functions by checking fitness of the model's on data training and for the complexity of the model, it adds a penalty. The result that is desirable is the lowest possible AIC, which depicts model fitness. The model is evaluated by AIC on the basis of maximum likelihood estimation. This is expressed as:

$$AIC = 2K - 2(log - likelihood) \quad (3)$$

AIC uses the maximum likelihood estimation of the model (log-likelihood) to measure fitness. Log-likelihood is a measure of how likely one is to see their observed data, given a model.

Maximum likelihood indicates the best fitness of the model to the dataset. AIC uses computational convenience instead of natural log of the likelihood.

Methodology

Data Collection and Description. The most widely used methods to determine toxicological risk in an aqueous medium are inhibition assays which use Vibrio fischeri (formerly known as Photobacterium Phosphoreum), a marine gram-negative bacterium[39]. Many different luminescence inhibition tests involving Vibrio fischeri have been developed for the analysis of aqueous samples [10]. The toxicity data used in this study are log (EC50) where EC50 is in $\mu molL^{-1}$. The EC50 is the concentration of a compound at which 50% of its maximal effect is observed [16]. These totoxicity values which span over five log units are reported for 74 ILs including thirty-five heterocyclic cations and eighteen organic and inorganic anions available in literature [27]. The descriptors used in this study were developed by Paternò et al. [27] using principal component analysis (PCA) to compact VolSurf+ derived in silico molecular properties into nine principal properties. The variables used are 128 cationic and 38 anioinc variables available in VolSurf+. Five principal components which described 77.5% of variance are considered significant to the PCA model and chosen as cationic PPs. On the other hand, four principal components explained 73.5% of the variance were chosen as anionic PPs [27].

The first cationic PP decribes the solubility, size, flexibility and molecular weight of the cation. The second describes cation's interaction with water and the hydrophobic volume of the cation. The third cationic PP evidences the difference between more amphiphilic ILs from those with a higher hydrophobic character. The fourth cationic PP includes descriptors such as skin permeability, hydophilic/hydrophobic ratio, permeability into CACO2 cells blood-brain barrier permeation and ability to form hydrogen bond as donor. The fifth cationic PP is mainly required to discriminate hydrogen bond donor descriptors exhibiting high PPC5 values from hydrogen bond acceptors exhibiting very high negative PPC5 values. On the other hand, first and second anionic PPs encompasses descriptors related to hydrophobicity/hydrophilicity balance, and critical packing. The first aninic PP distinguishes anions based on their hydrophobicity/hydrophilicity balance. The third and fourth anionic PPs descriptors related to the size/shape of the anions and to the anion's ability to form hydrogen bond as donor or acceptor. The third anionic PP differentiate anions with hydrophilicity due to polarizability from those with hydrophilicity due to the anion ability to form hydrogen bond as donor or acceptor. Detailed explanation of the various PPs could be found in [27].

Data Analysis and Visualization. For the entire workflow in this study, python 3.7 programming language of Jupyter notebook was used. The data was subjected to a series of techniques to process, clean and select variables that are relevant to the machine learning models. The relationship between the input (principal properties) and output (Vibrio fischeri toxicities) data was explored as a summary statistic of the input and output data are given in Table 1. A pair plot gave the pictorial relationship between the descriptors and target data as indicated in Fig. 1. Linear regression will not be an appropriate model for this study as indicated by the nonlinear distribution of the data. There is the need to remove colinear features which could be achieved by using advanced feature selection methods. This is because the distribution shows presence of multicollinearity that is likely to cause uncertainty in any model for machine learning. To quantify the degree of correlation among the descriptors spearman rho correlation covariance matrix was calculated as shown in Fig. 2. This analysis useful in the determination of the distribution of data and helps in telling their expected behaviour.

Table 1. Summary statistics of input and output data

	Vibrio_f	PP1C	PP2C	PP3C	PP4C	PP5C	PP1A	PP2A	PP3A	PP4A
count	74	520	520	520	520	520	520	520	520	520
mean	3.18	0.84	-1.67	0.20	0.03	-0.05	0.01	-2.13	-0.65	-0.08
std	1.36	5.50	4.99	3.59	3.01	1.87	2.40	3.27	2.49	2.31
min	-0.18	-21.63	-8.24	-9.65	-8.83	-9.43	-13.16	-5.73	-5.87	-3.83
25%	2.50	-2.70	-5.91	-2.39	-1.31	-1.17	-1.19	-5.52	-2.31	-1.25
50%	3.24	0.58	-3.31	1.33	0.51	-0.19	-0.57	-2.15	-1.63	-1.13
75%	4.14	4.71	2.25	2.76	1.70	0.86	1.35	-0.06	1.04	2.10
max	6.10	14.74	15.45	6.77	5.90	5.63	4.81	7.57	6.72	4.62

Fig. 1. Pair plot distribution of PPs and Vibrio fischeri toxicity

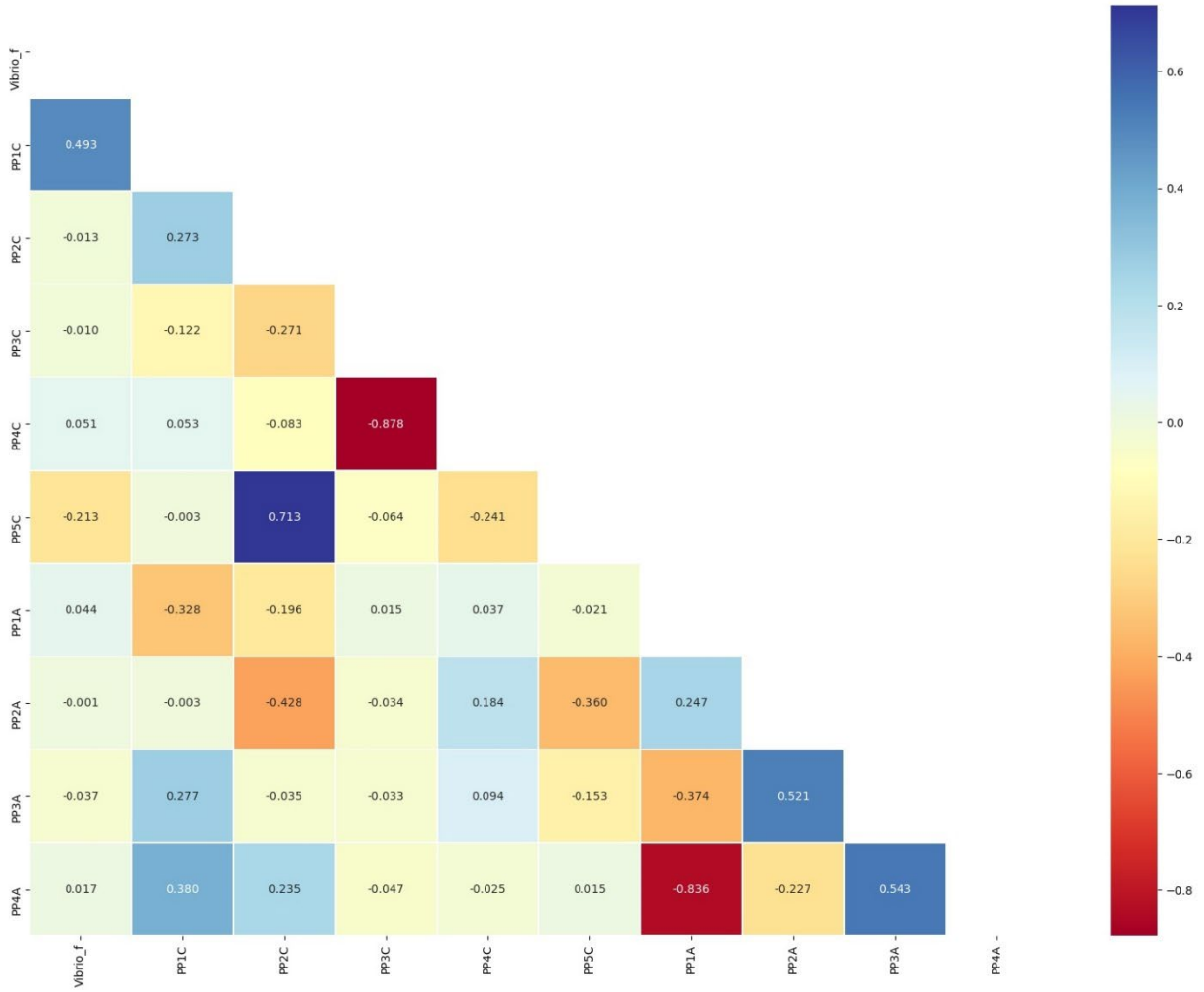

Fig. 2. Plot of the Spearman rho correlation covariance matrix for all the data ranging from -1 to 1.

The various ML models were used to estimate *Vibrio fischeri* toxicities from all the PPs. A section of data that was not used in the learning process (secondary section) was used to cross validate the model. Cross validation was to check model accuracy in predicting the whole dataset. The technique used here for cross validation was holdout which works by using a new subset of data (i.e., data not used in model training) on the model. This technique finds ways to prevent occurrence of over or under-fitting, offer an insight into bias-variance trade-offs and minimize estimation errors on unobserved data [40]. The held-out data was later utilized in the validation of the accuracy of the prediction model. The data was split into two, i.e., 85% (62 data points) was used as training data and 15% (12 data points) as test data. This is to ensure that validation of the prediction accuracy of the model is done with another set of data set aside from model training. It is imperative to note that the focus is not the model accuracy but the prediction error. The variance and bias of the model is better understood by analyzing the errors. The bias is the error rate, and the selection of input data is the most influencing parameter on a model's bias. On the other hand, the variance is the performance deterioration of the of the model in terms of accuracy when used on the test data compared to when used on the training data.

Results and Discussion

Model Performances Evaluation. Model performance for various models is done using the test data which was held out in training these models. The training and test scores for the various models are presented in Table 2. As seen from Table 2, the training accuracy for each model is very high as expected. This is because the model tries to replicate data that has already been seen. Nevertheless, a high training accuracy does not necessary depict that the model is a good one as it could be a non-generalized model produced due to overfitting where inherent noise is captured by the model. Another scenario could be an out-of-sample accuracy which is the correct prediction percentage that the model makes on data not used in training it. It is imperative for models to have an out-of-sample accuracy that is high as they are meant to make correct predictions on unknown data. All the models had low out-of-sample accuracy due to small data set with high dimension. However, decision tree proved to have the most accurate out-of-sample prediction among the models trained.

Table 2. Correlation coefficient score for train and test data

Models	Train Score	Test Score
Decision Tree	1	0.613051016
Random Forest	0.930593458	0.571406118
Extra tree Regression	1	0.474419201
Gradient Boosting Regression	0.981733285	0.571829785
XGBoost	0.999996143	0.444828751

In terms of probabilistic ranking of models' performances, the likelihood of the models to reduce loss of information is determined by AIC. A more parsimonious model usually has lower AIC values compared to other models [41]. This makes the chosen model better for prediction. As shown in Fig. 3 the decision tree is best in predicting *Vibrio fischeri* toxicity among the models trained in this study. The AIC assesses the convergence of the models fit to actual data by computing its relative quality. Nevertheless, the model's tendency to overfit or underfit still exists. Therefore, the precision, consistency, and accuracy of the models' performance were also evaluated.

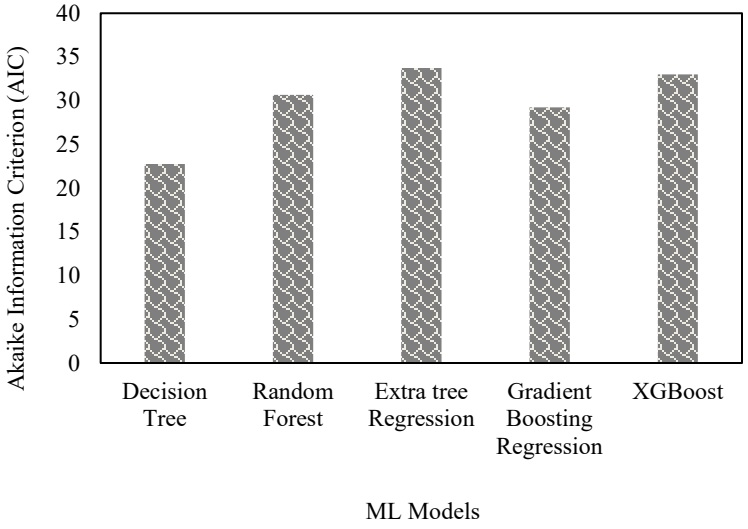

Fig. 3. Comparison of AIC for all models on test data

The models are further assessed on their improvement in accuracy (i.e., reduction in error). The models deployed in this study have varying performances that are very much due to their statistical

and theoretical framework. Cross-validation analysis is performed to examine the model performance in terms of precision, accuracy, and consistency in predicting unknown toxicities. This analysis is performed using the RMSE and MAE as shown in Fig. 4. The RMSE of all the models indicated that the decision tree model is the most consistent in predicting *Vibrio fischeri* toxicity. Nevertheless, the MAE results showed that the most precise model for *Vibrio fischeri* toxicity prediction is RF model. Therefore, as regards fitness, consistency and accuracy of the model, decision tree model outperforms all the other models deployed in this study. Comparing the performance of the models based on these error matrices and the AIC, the decision tree proves to be the most robust and accurate model trained on this data set.

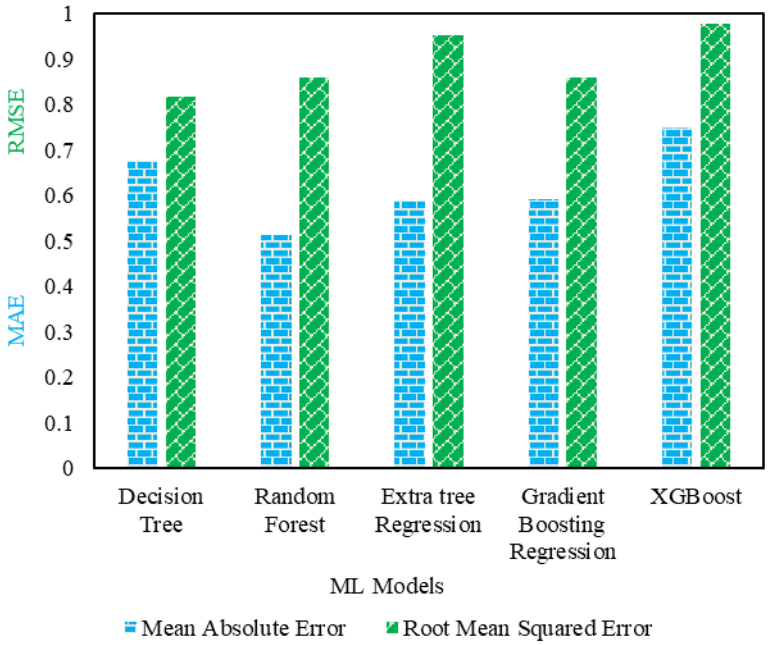

Fig. 4. Comparison of prediction errors for all models

It is imperative to point out that variation in model performance could also be attributed to the data set given. Some models work better on large data set while others work better on smaller data set. Generally, XGBoost model tend to perform best among the models deployed in this study however, the decision tree proved robust owing to the small data set used in this study. Another inherent property with XGBoost model is the tendency to convey the significance of the descriptors to the output based on F-score as portrayed in Fig. 5. From Fig. 5, PP1C is the most important followed by PP1A then PP3C. Paternò et al. [16] in their prediction of *Vibrio fischeri* toxicity also used a plot of the Variable Importance for Projection (VIP) to show the significance of the various PPs to *Vibrio fischeri* toxicity. PP1C was the most pertinent descriptor followed by PP2C then PP5C (which had high error) then PP3A. The other five descriptors were less important. Though different variations of the descriptor importance have been obtained, the PP1C prove to be the most important descriptor in both cases. This shows how the cationic part of IL liquids are important to their toxicity. The size, flexibility, solubility and molecular weight of the cation is information represented by PP1C. In this study, the XGBoost also shows how the anionic part of IL is also important to their toxicity and the need to consider both counterparts in toxicity prediction.

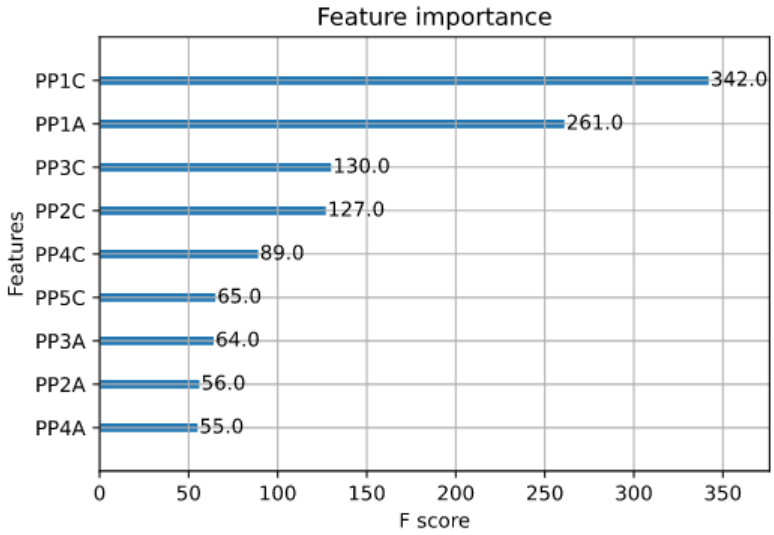

Fig. 5. Feature importance of all descriptors to Vibrio fischeri toxicity prediction

Further evaluation of model performances is conducted by comparing the predicted toxicities using descriptors of test data to the actual test toxicities. Figs. 6 and 7 show joint plots of correlation plot of predicted toxicities versus actual toxicities together with the distributions of the toxicities for decision tree and XGBoost. The shaded area around the correlation line shows the deviation of the predicted values. The wider the shaded area, the more deviated the predicted toxicities are from the actual toxicity values. By comparing Fig. 6 and 7, the decision tree proved to have better predictive power over the XGBoost in predicting *Vibrio fischeri* toxicity. Nevertheless, the predicted values for both models had different distribution compared to the actual values. This shows how less consistent even the decision tree is in predicting *Vibrio fischeri* toxicity.

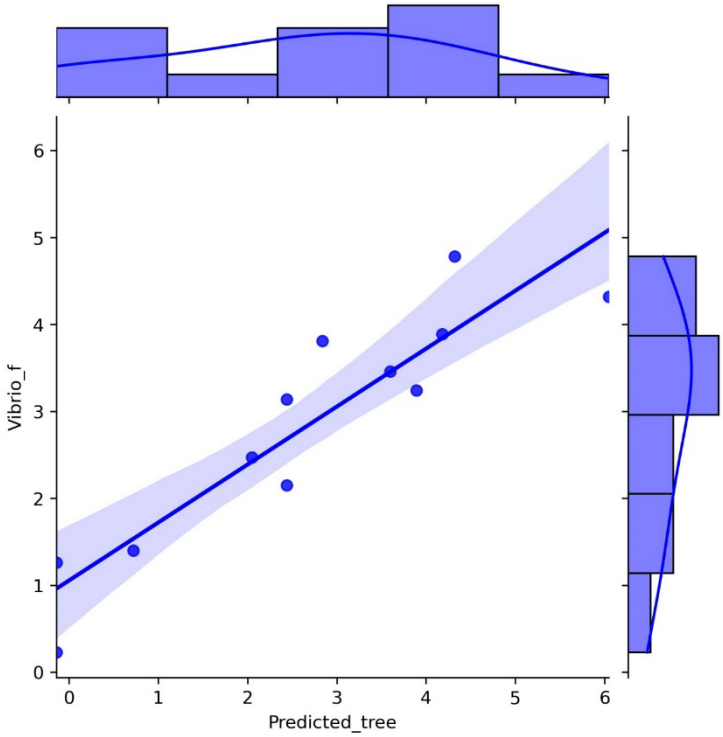

Fig. 6. Joint plot of Actual against predicted toxicities by decision tree

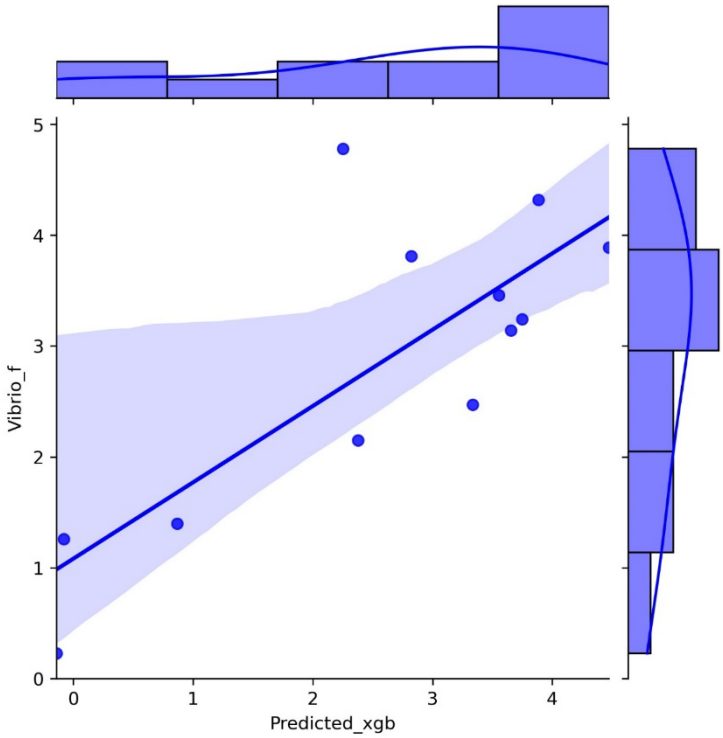

Fig. 7. Joint plot of Actual against predicted toxicities by XGBoost

Sensitivity Analysis. The kernel density estimates of *Vibrio fischeri* toxicity by decision tree and XGBoost are presented in Fig. 8 and 9 respectively. The red curve represents actual values while the blue one is predicted values. In this case, the XGBoost had its predicted values closer to the actual values than the decision tree. The XGBoost showed better predictive power in higher values than lower values while the values predicted by the decision tree showed wider distribution with lower mode. The toxicity of *Vibrio fischeri* is expressed as EC50 which signifies the compound's concentration at which 50% of its maximal effect is observed, hence the lower this value is, the more toxic the IL. Therefore, the XGBoost was biased toward predicting lower toxicities more accurately than higher toxicities. The decision tree showed consistency in its prediction irrespective of the error.

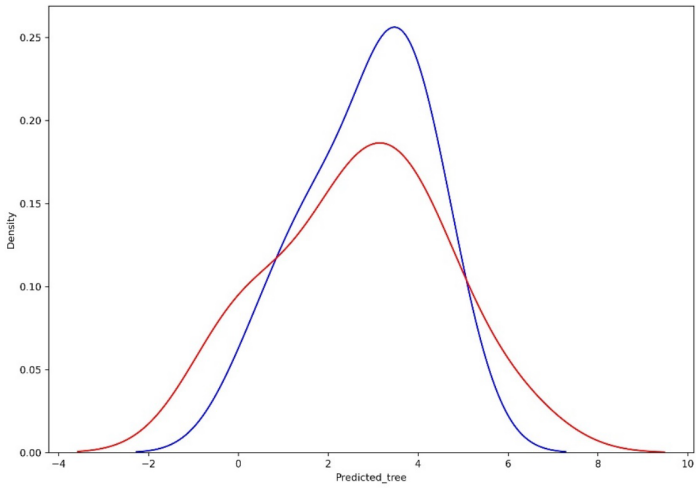

Fig. 8. Kernel density estimation: The closeness of predicted to actual values for Decision tree Vibrio fischeri toxicity prediction

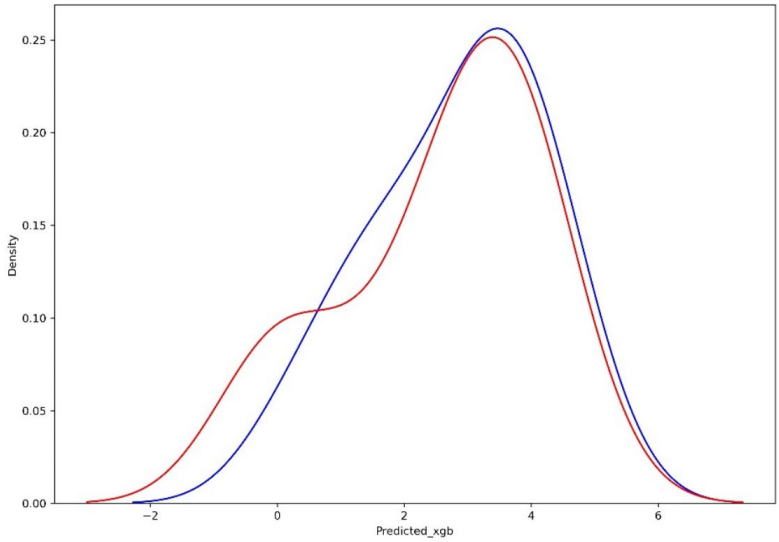

Fig. 9. Kernel density estimation: The proximity of predicted values to actual values for the prediction of Vibrio fischeri toxicity using the XGBoost model.

Conclusion

In this study, ML models have been trained and tested for prediction of *Vibrio fischeri* toxicities of IL using nine in silico descriptors developed by Paternò et al. [16], [27]. Comparing their training and test scores, most of the model showed evidence of overfitting. Nevertheless, the decision tree model proved to be the most consistent and accurate model among the models deployed in this study. Model performance based on RMSE showed that the RF model was the most precise. Other models are generally deemed more accurate than the decision tree, therefore, the observed performance could be attributed to the small data set size used. The inherent attribute of the XGBoost showed that both PP1C and PP1A are very important to *Vibrio fischeri* toxicity prediction. This proved that both the cationic and anionic counterparts of IL are important in determining their level of toxicities. Further ML techniques could be deployed to make the models more robust including feature selection and optimization. This study proves that QSPR studies on toxicity prediction of new ILs could be improved via ML application.

Acknowledgement

The authors would like to acknowledge the FRGS Grant (015MA0-126), Universiti Teknologi PETRONAS for the financial assistance.

Declaration of conflict of interest

On behalf of all the co-authors, the corresponding author states that there is no conflict of interest.

References

[1] F. D'Anna, S. Marullo, P. Vitale, and R. Noto, "Synthesis of aryl azides: A probe reaction to study the synergetic action of ultrasounds and ionic liquids," *Ultrason. Sonochem.*, vol. 19, no. 1, pp. 136–142, 2012, https://doi.org/10.1016/j.ultsonch.2011.06.010.

[2] N. Taccardi *et al.*, "Gallium-rich Pd-Ga phases as supported liquid metal catalysts," *Nat. Chem.*, vol. 9, no. 9, pp. 862–867, 2017, https://doi.org/10.1038/NCHEM.2822.

[3] M. Tariq, D. Rooney, E. Othman, S. Aparicio, M. Atilhan, and M. Khraisheh, "Gas hydrate inhibition: A review of the role of ionic liquids," *Ind. Eng. Chem. Res.*, vol. 53, no. 46, pp. 17855–17868, 2014, https://doi.org/10.1021/ie503559k.

[4] M. El-Harbawi, "Toxicity Measurement of Imidazolium Ionic Liquids Using Acute

Toxicity Test," *Procedia Chem.*, vol. 9, no. December 2014, pp. 40–52, 2014, https://doi.org/10.1016/j.proche.2014.05.006.

[5] C. B. Bavoh, O. Nashed, A. N. Rehman, N. A. A. B. Othaman, B. Lal, and K. M. Sabil, "Ionic Liquids as Gas Hydrate Thermodynamic Inhibitors," *Ind. Eng. Chem. Res.*, vol. 60, no. 44, pp. 15835–15873, 2021, https://doi.org/10.1021/acs.iecr.1c01401.

[6] C. B. Bavoh, T. N. Ofei, and B. Lal, "Investigating the Potential Cuttings Transport Behavior of Ionic Liquids in Drilling Mud in the Presence of sII Hydrates," *Energy and Fuels*, vol. 34, no. 3, pp. 2903–2915, 2020, https://doi.org/10.1021/acs.energyfuels.9b04088.

[7] E. E. L. Tanner, R. R. Hawker, H. M. Yau, A. K. Croft, and J. B. Harper, "Probing the importance of ionic liquid structure: A general ionic liquid effect on an SNAr process," *Org. Biomol. Chem.*, vol. 11, no. 43, pp. 7516–7521, 2013, https://doi.org/10.1039/c3ob41634h.

[8] J. Scholz *et al.*, "Ethylene to 2-butene in a continuous gas phase reaction using silp-type cationic nickel catalysts," *ChemCatChem*, vol. 6, no. 1, pp. 162–169, 2014, https://doi.org/10.1002/cctc.201300636.

[9] R. Mancuso, C. S. Pomelli, C. Chiappe, R. C. Larock, and B. Gabriele, "A recyclable and base-free method for the synthesis of 3-iodothiophenes by the iodoheterocyclisation of 1-mercapto-3-alkyn-2-ols in ionic liquids," *Org. Biomol. Chem.*, vol. 12, no. 4, pp. 651–659, 2014, https://doi.org/10.1039/c3ob41928b.

[10] P. M. Ventura, C. S. Marques, A. A. Rosatella, C. A. M. Afonso, and F. Gonc, "Ecotoxicology and Environmental Safety Toxicity assessment of various ionic liquid families towards Vibrio fischeri marine bacteria," vol. 76, pp. 162–168, 2012, https://doi.org/10.1016/j.ecoenv.2011.10.006.

[11] S. P. F. Costa, P. C. A. G. Pinto, R. A. S. Lapa, and M. L. M. F. S. Saraiva, "Toxicity assessment of ionic liquids with Vibrio fischeri: An alternative fully automated methodology," *J. Hazard. Mater.*, vol. 284, pp. 136–142, 2014, https://doi.org/10.1016/j.jhazmat.2014.10.049.

[12] R. N. Das, T. E. Sintra, J. A. P. Coutinho, S. P. M. Ventura, K. Roy, and P. L. A. Popelier, "Development of predictive QSAR models for: Vibrio fischeri toxicity of ionic liquids and their true external and experimental validation tests," *Toxicol. Res. (Camb).*, vol. 5, no. 5, pp. 1388–1399, 2016, https://doi.org/10.1039/c6tx00180g.

[13] F. G. Doherty, "A review of the Microtox® toxicity test system for assessing the toxicity of sediments and soils," *Water Qual. Res. J.*, vol. 36, no. 3, pp. 475–518, 2001.

[14] M. Ishaq, D. Zaini, A. Mohd, and M. Moniruzzaman, "Framework for Ecotoxicological Risk Assessment of Ionic Liquids," *Procedia Eng.*, vol. 148, pp. 1141–1148, 2016, https://doi.org/10.1016/j.proeng.2016.06.567.

[15] M. I. Khan, D. Zaini, and A. M. Shariff, "Estimation of safe environmental concentrations of ionic liquids towards bacteria by chemical toxicity distribution method Estimation of safe environmental concentrations of ionic liquids towards bacteria by chemical toxicity distribution method," 2018, https://doi.org/10.1088/1757-899X/458/1/012038.

[16] A. Paterno', S. Scire, and G. Musumarra, "A QSPR approach to the ecotoxicity of ionic liquids (Vibrio fischeri) using VolSurf principal properties," *Toxicol. Res. (Camb).*, vol. 5, no. 4, pp. 1090–1096, 2016.

[17] A. Romero, A. Santos, J. Tojo, and A. Rodr, "Toxicity and biodegradability of imidazolium ionic liquids," vol. 151, pp. 268–273, 2008,

https://doi.org/10.1016/j.jhazmat.2007.10.079.

[18] L. Maltby, "Small-Scale Freshwater Toxicity Investigations: Volume 1 ? Toxicity Test Methods," *Freshw. Biol.*, vol. 52, no. 1, pp. 198–198, 2007, https://doi.org/10.1111/j.1365-2427.2006.01662.x.

[19] S. M. Steinberg, E. J. Poziomek, W. H. Engelmann, and K. R. Rogers, "A review of environmental applications of bioluminescence measurements," *Chemosphere*, vol. 30, no. 11, pp. 2155–2197, 1995, https://doi.org/10.1016/0045-6535(95)00087-O.

[20] S. P. F. Costa, V. D. Justina, K. Bica, M. Vasiloiu, P. C. A. G. Pinto, and M. L. M. F. S. Saraiva, "Automated evaluation of pharmaceutically active ionic liquids'(eco) toxicity through the inhibition of human carboxylesterase and Vibrio fischeri," *J. Hazard. Mater.*, vol. 265, pp. 133–141, 2014.

[21] M. Jafari, M. H. Keshavarz, and H. Salek, "A simple method for assessing chemical toxicity of ionic liquids on Vibrio fischeri through the structure of cations with specific anions," *Ecotoxicol. Environ. Saf.*, vol. 182, p. 109429, 2019.

[22] M. G. Montalbán, J. M. Hidalgo, M. Collado-González, F. G. D. Baños, and G. Víllora, "Assessing chemical toxicity of ionic liquids on Vibrio fischeri: Correlation with structure and composition," *Chemosphere*, vol. 155, pp. 405–414, 2016.

[23] K. M. Docherty and C. F. Kulpa, "Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids," pp. 185–189, 2005, https://doi.org/10.1039/b419172b.

[24] B. Peric, J. Sierra, E. Martí, R. Cruañas, and M. A. Garau, "Quantitative structure–activity relationship (QSAR) prediction of (eco) toxicity of short aliphatic protic ionic liquids," *Ecotoxicol. Environ. Saf.*, vol. 115, pp. 257–262, 2015.

[25] F. Yan, Q. Shang, S. Xia, Q. Wang, and P. Ma, "Topological study on the toxicity of ionic liquids on Vibrio fischeri by the quantitative structure-activity relationship method," *J. Hazard. Mater.*, vol. 286, pp. 410–415, 2015, https://doi.org/10.1016/j.jhazmat.2015.01.016.

[26] D. J. Couling, R. J. Bernot, K. M. Docherty, J. N. K. Dixon, and E. J. Maginn, "Assessing the factors responsible for ionic liquid toxicity to aquatic organisms via quantitative structure–property relationship modeling," *Green Chem.*, vol. 8, no. 1, pp. 82–90, 2006, https://doi.org/10.1039/b511333d.

[27] A. Paternò *et al.*, "Cyto- and enzyme toxicities of ionic liquids modelled on the basis of VolSurf+ descriptors and their principal properties," *SAR QSAR Environ. Res.*, vol. 27, no. 3, pp. 221–244, 2016, https://doi.org/10.1080/1062936X.2016.1156571.

[28] A. Paternò, G. Bocci, L. Goracci, G. Musumarra, and S. Scirè, "Modelling the aquatic toxicity of ionic liquids by means of VolSurf in silico descriptors," *SAR QSAR Environ. Res.*, vol. 27, no. 1, pp. 1–15, 2016.

[29] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[30] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *Int. J. Inf. Decis. Sci.*, vol. 12, no. 3, pp. 246–269, 2020.

[31] K. Mittal, D. Khanduja, and P. C. Tewari, "An insight into 'Decision Tree Analysis'"," *World Wide J. Multidiscip. Res. Dev.*, vol. 3, no. 12, pp. 111–115, 2017.

[32] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Adv. Sp. Res.*, vol. 41, no. 12, pp. 1955–1959, 2008.

[33] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[34] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, “Hough forests for object detection, tracking, and action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, 2011.

[35] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann. Stat.*, pp. 1189–1232, 2001.

[36] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[37] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, M. Y. Taki, and B. N. Tackie-Otoo, “Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions,” *J. Pet. Sci. Eng.*, p. 109244, 2021.

[38] H. Bozdogan, “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions,” *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[39] S. Parvez, C. Venkataraman, and S. Mukherji, “A review on advantages of implementing luminescence inhibition test (Vibrio fischeri) for acute toxicity prediction of chemicals,” *Environ. Int.*, vol. 32, no. 2, pp. 265–268, 2006, https://doi.org/10.1016/j.envint.2005.08.022.

[40] W. Ertel, *Introduction to artificial intelligence*. Springer, 2018.

[41] D. A. Otchere, T. O. A. Ganat, R. Gholami, and M. Lawal, “A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction,” *J. Nat. Gas Sci. Eng.*, vol. 91, no. April, p. 103962, 2021, https://doi.org/10.1016/j.jngse.2021.103962.