Arabian Journal of Chemistry (2022) **15**, 104228

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Novel and accurate mathematical simulation of various models for accurate prediction of surface tension parameters through ionic liquids

Rami J. Obaid [a], Hossam Kotb [b], Amal M. Alsubaiyel [c], Jalal Uddin [d], Mohd Sani Sarjad [e], Md. Lutfor Rahman [e,*], Saleh A. Ahmed [a,f,*]

[a] *Department of Chemistry, Faculty of Applied Science, Umm-Al-Qura University, 21955 Makkah, Saudi Arabia*
[b] *Department of Electrical Power and Machines, Faculty of Engineering, Alexandria University, Alexandria, Egypt*
[c] *Department of Pharmaceutics, College of Pharmacy, Qassim University, Buraidah 52571, Saudi Arabia*
[d] *Department of Pharmaceutical Chemistry, College of Pharmacy, King Khalid University, Asir 61421, Saudi Arabia*
[e] *Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu 88400, Sabah, Malaysia*
[f] *Department of Chemistry, Faculty of Science, Assiut University, 71516 Assiut, Egypt*

Received 16 June 2022; accepted 30 August 2022
Available online 5 September 2022

KEYWORDS

Ionic liquids;
ADABOOST-GPR model;
Surface tension;
Simulation

Abstract Ionic Liquids (ILs) as a novel class of liquid solvent simultaneously carry the positive characteristics of both molten salts and organic liquids. Remarkable positive properties of ILs have such as low vapor pressure and excellent permittivity have encouraged the motivation of researchers to use them in various applications over the last decade. Surface tension is an important physico-chemical property of ILs, which its experimental-based measurement has been done by various researchers. Despite great precision, some major shortcomings such as high cost and health-related problems caused the researchers to develop mathematical models based on artificial intelligence (AI) approach to predict surface tension theoretically. In this research, the surface tension of two novel ILs (bis [(trifluoromethyl) sulfonyl] imide and 1,3-nonylimidazolium bis [(trifluoromethyl) sulfonyl] imide) were predicted using three predictive models. The available dataset contains 45 input features, which is relatively high in dimension. We decided to use AdaBoost with different base models, including Gaussian Process Regression (GPR), support vector regression (SVR), and decision tree (DT). Also, for feature selection and hyper-parameter tuning, a genetic algorithm (GA) search is used. The final R^2-score for boosted DT, boosted GPR, and boosted SVR is 0.849, 0.981, and 0.944, respectively. Also, with the MAPE metric, boosted GPR has an

* Corresponding authors.
E-mail addresses: lotfor@ums.edu.my (Md. Lutfor Rahman), saahmed@uqu.edu.sa (S.A. Ahmed).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.arabjc.2022.104228

error rate of 1.73E-02, boosted SVR has an error rate of 2.35E-02, and it is 3.36E-02 for boosted DT. So, the ADABOOST-GPR model was considered as the primary model for the research.

1. Introduction

The emergence of paramount attentions towards the application of eco-friendly materials in different industries, ILs have found their high place thanks to their brilliant advantages compared to conventional toxic solvents (Járvás et al., 2018; Wang et al., 2014). ILs belongs to a group of chemical salts formed by an asymmetric organic cation and a symmetric inorganic anion. Their great potential of liquefaction at the temperatures near/below room temperature, makes them ideal as a potent solvent for polar organic molecules, inorganic salts and transition metal catalyzed reactions (Fehér et al., 2012; Shojaeian and Asadizadeh, 2020).

In current decades, ILs have been introduced as promising options for disparate industrial-based applications including chemical process, reaction engineering, metals recovery and membrane-based separation owing to their indisputable positive points such as excellent permittivity, acceptable solubility and thermal stability (Zia ul Mustafa et al., 2019; Shang et al., 2017; Kianfar and Mafi, 2021; Lal et al., 2021; Mahandra et al., 2021). True recognition of physicochemical properties of ILs like melting point, vapor pressure and surface tension is an mandatory process in the development of an industrial system.

Surface tension of an IL is a momentous factor in chemical industries due to its noteworthy impacts on the calculation of heat and mass transfers for the accurate design/development of disparate processes like gas absorption and distillation (Hashemkhani et al., 2015; Shojaeian, 2018). Despite the great importance of experimental investigations for measuring ILs' surface tension, their remarkable drawbacks such as health problems and high process cost have motivated the scientists to try harder to apply different mathematical/predictive procedures to compute different properties of ILs (Esmaeili and Hashemipour, 2021; Mjalli et al., 2014; Mirkhani et al., 2013). Apart from the practical advantages, development precise mathematical models may facilitate the researches on the theoretical interconnection of ILs structures and physical properties (Gardas and Coutinho, 2008).

Machine learning models to uncover useful information from experimental data are one of the most significant advancements that have touched numerous scientific domains nowadays. This fact has had an impact on the majority of experimental sciences. Ensemble models are an important class of machine learning techniques. The generalizations of boosting, bagging, random subspace, and stacking Ensemble models are many ensemble kinds based on ensemble models' characteristics of requiring less processing work and providing more accuracy in certain applications. (Zhou, 2019; Dietterich, 2000; Kadavi et al., 2018; Goodfellow et al., 2016). In this work, we applied boosting models, especially Adaboost, on top of SVR, GPR, and DT base models.

A decision tree regressor (DT) is a simple, understandable, and effective method to many estimate issues. The decision tree algorithm's main premise is to divide a huge issue into several smaller subproblems (Divide-and-conquer), which may result in an easier-to-understand answer (Xu et al., 2005; Song and Ying, 2015).

Support vector regressor (SVR) and Gaussian process regressor (GPR) are two other base learners. The first is based on the concept of locating a hyperplane that optimally separates inputs into different zones. The second is a nonparametric, Bayesian regression method that is making waves in the area of machine learning. The ability to cope with small datasets and offer uncertainty metrics on predictions are only two of the benefits of Gaussian process regression (An et al., 1964; Wang et al., 2020; Wilson et al., 2011; Shi and Choi, 2011; Kecman, 2005; Moosaei et al., 2021).

2. Data Set

In this study, a massive dataset of 1042 data rows from 69 ionic liquids was used. At constant pressure, surface tension and temperature are measured at intervals of [18.5, 70.3] (mN/m) and [268.29, 532.4] (K) The inputs were temperature and chemical structure, and the output was surface tension (Mousavi et al., 2021).

3. Methodology

3.1. Base Models

Decision tree Regressor (Regression Tree) is one of weak predictors employed in this study and boosted using Adaboost. Decision Tree gives a class of questions through a set of properties like 'is greater' or 'is equal' with the provided *True* or *False* responses, another query will be met to respond. This operation is repeated until no more inquiries are obtained. The information is continually split into dual ingredients, which allows the Decision Tree to be created. A randomness metric like entropy has applied to assess the divisions for all attributes (Mathuria, 2013; Sakar and Mammone, 1993).

Derived from statistical learning theory, SVR is a sophisticated learning algorithm. For Vapnik, this method was pioneered (Vapnik, 1999). SVR has been trained to recognize the dependency relationship between a collection of goals $t = \{t_1, t_2, \cdots, t_n\}$ specified on R and inputs $x = \{x_1, x_2, \cdots, x_n\}$ that $x_i \in R^d$, where n is count of instances in the dataset. Linear regression is applicable to solve this problem because the problem has been transformed into multidimensional feature. The following equation may be used to illustrate this concept (Dargahi-Zarandi et al., 2020):

$$f(x) = w.\phi(x) + b$$

In this example, the mapping function $\phi(x)$ can turn an input vector into something else. b and w represent the bias and weight axes (Amar et al., 2020).

To calculate w and b, the so-called "regularized risk function" is used to incorporate the model's complexity and related experimental error into a regression-purpose optimization problem. In other words, there is a problem in the system (Keane et al., 2008):

$$minimize \frac{1}{2}\|w\|^2 + C \times \sum_{k=1}^{n} \left(\xi_k^- + \xi_k^+\right)$$

$$s.t.\{t_k - (w.\phi(x_k) + b) \le \varepsilon + \xi_k^+ (w.\phi(x_k) + b) - t_k$$
$$\le \varepsilon + \xi_k^- \xi_k^-, \xi_k^+ \ge 0, i \in \{1, 2, \cdots, n\}$$

Where $\sum_{k=1}^{n}(\xi_k^- + \xi_k^+)$ stands for empirical error and $\|w\|^2$ reflects function flatness. The inclusion of a penalty constant, C, regulates model complexity and related empirical inaccuracy to some extent. Also, ε is error tolerance; and $\xi_k^+ and \xi_k^-$ are positive values that reflect upper and lower excess deviations.

Lagrange multipliers convert the above-mentioned constrained optimization issue within a double carrier. The resulting Lagrangian has eliminated. The resolution phases have already been detailed (Keane et al., 2008). The resultant response is:

$$f(x) = \sum_{k=1}^{n}(\alpha_k - \alpha_k^*)K(x_k, x_l) + b$$

$K(x_k, x_l)$ denotes the kernel function and α_k *and* α_k^* shows the Lagrange multipliers, then $0 \le \alpha_k$ *and* $\alpha_k^* \le C$.

The other employed base model is Gaussian process regression (GPR). GPR does not require declaration of a fitting function (Quinonero-Candela and Rasmussen, 2005; Jiang et al., 2021).

y is demonstrated as $f(\boldsymbol{x})$ for a set of n-dimensional instances $D = \{(\boldsymbol{x}_i, y_i)|i = 1, 2, \cdots, n\}$, $\boldsymbol{x}_i \in R^d$ as input matrix $y_i \in R$ as output.

$$y = f(\boldsymbol{x})$$

Table 1 Outputs.

Models	MSE	R^2	MAPE
AdaBoost-DT	11.38	0.849	3.36E-02
AdaBoost-GPR	3.05	0.981	1.73E-02
AdaBoost-SVR	4.24	0.944	2.35E-02

GP can be defined through $f(x)$, as an implied function defined as a collection of random variables:

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), \mathbf{K})$$

In the above equation, K demonstrates any covariance showed by kernels and their corresponding input amounts, then $m(x)$ is the mean operator (Wu et al., 2020).

3.2. Adaboost

By bringing together numerous base estimators, it is feasible to develop an ensemble learner that outperforms an individual learner in generality and accuracy. Freund et al. (Freund and Schapire, 1997) proposed an ensemble strategy to develop the performance of individual learners through updating the weight of instances, then developed as the AdaBoost algorithm.

This strategy, as the name indicates, adaptively improves individual models, allowing them to tackle complex tasks. To deal with difficult problems, there are two approaches: simple models and advanced models. Further, simple models offer great generalization capabilities because of the simple structure. Despite their simplicity in real-world challenges, they are unable of tackling complicated problems owing to the substantial bias inherent in their structure.

Complicated models are more subject to be over fitted, and their application is more difficult in practice due to the problems in implementing them (Buitinck et al., 2013). These problems can be addressed using the AdaBoost strategy. Weak learner model is utilized as a foundation model in this strategy, and then other models are gradually combined to produce an

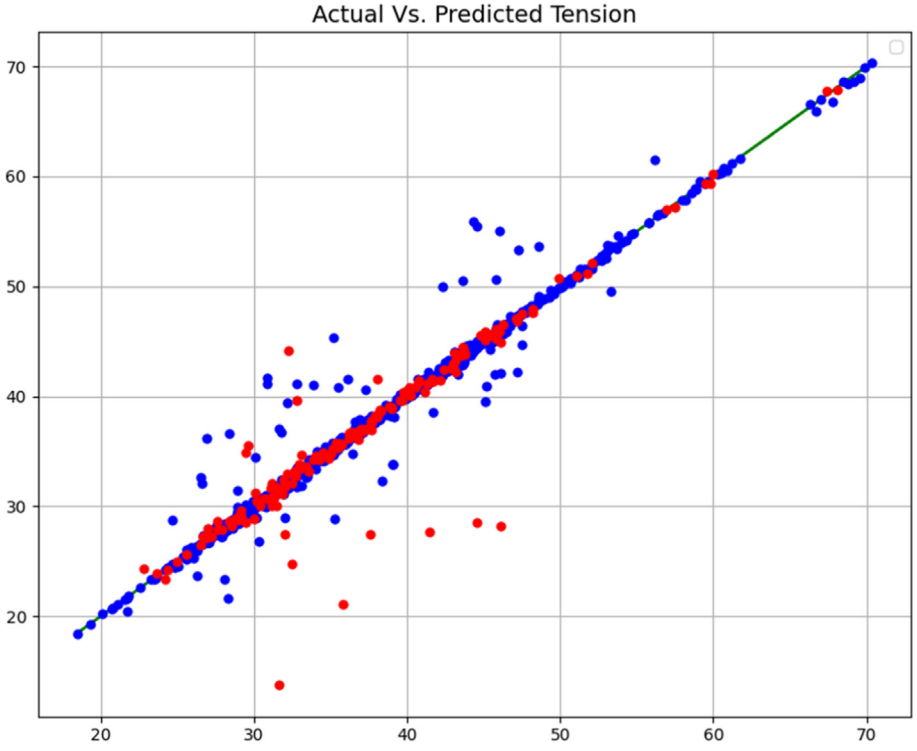

Fig. 1 Expected vs predicted (Boosted DT Model).

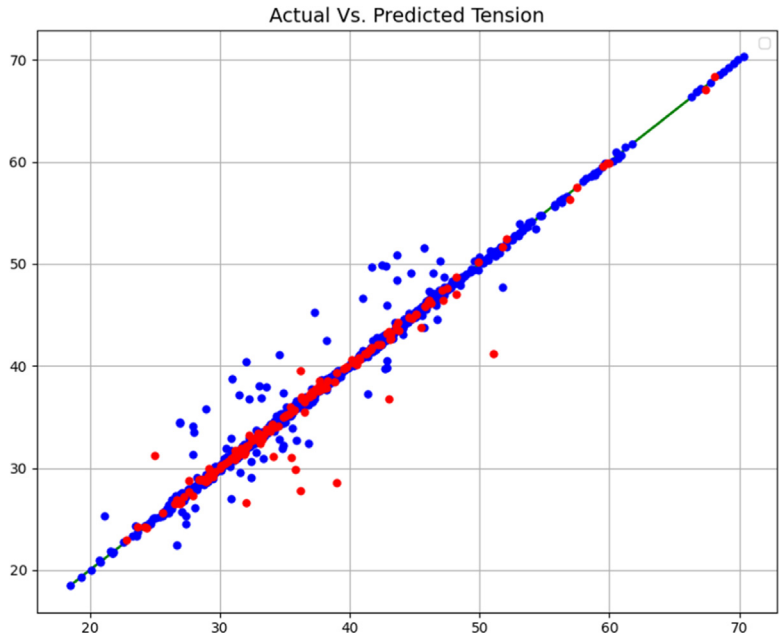

Fig. 2 Expected vs predicted (Boosted GPR Model).

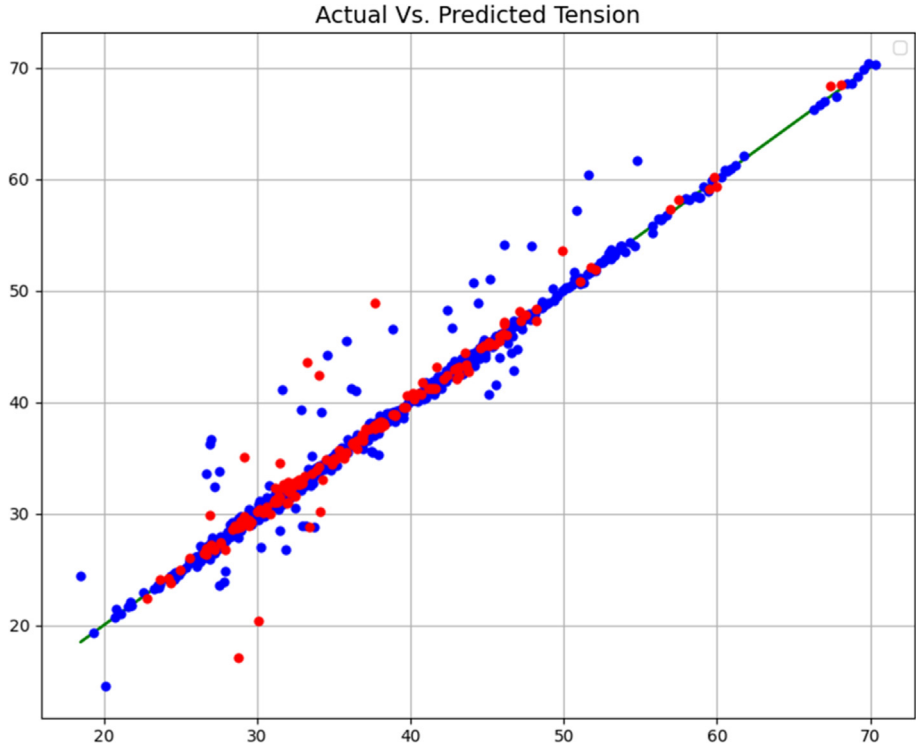

Fig. 3 Expected vs predicted (Boosted SVR Model).

strong system that can deal with complicated scenarios consistently and consistently (Lemaître et al., 2017).

4. Results

The above mentioned models were implemented and tuned to their hyper-parameters using genetic algorithm (GA). GA also applied for selection of features in order to gain better accuracy and generality. Then three metrics used to evaluate final models.

Without a doubt, R^2 is popular scale to evaluate the estimated outcome proficiency. That shows the efficiency of the projected discoverie's patterns which is correlated to the observed data's tendencies (Gouda et al., 2019).

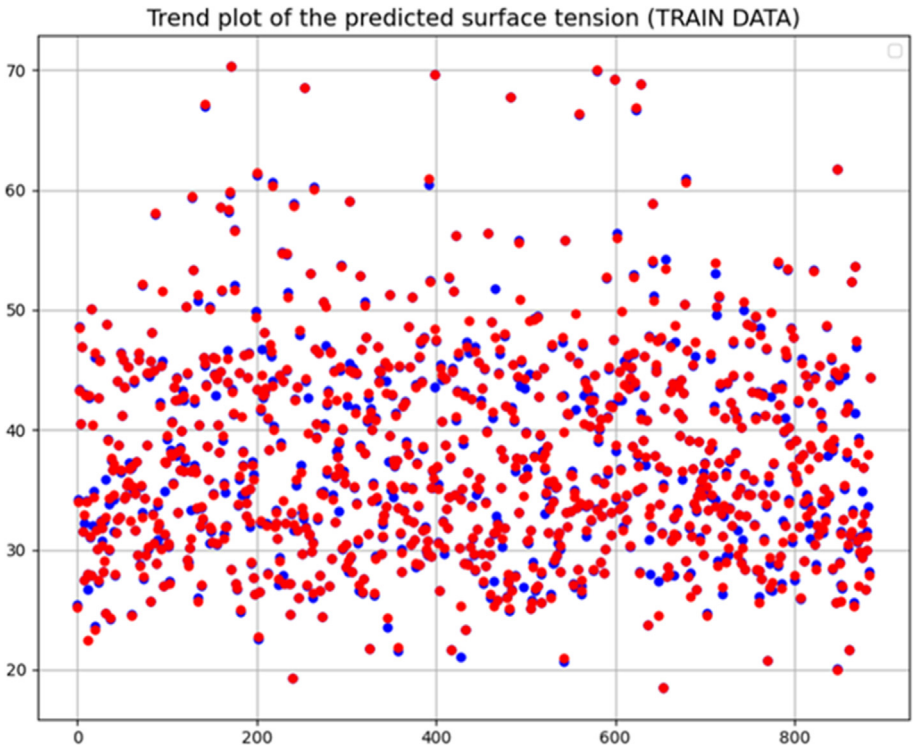

Fig. 4 Tendency of predicted surface tension – train data.

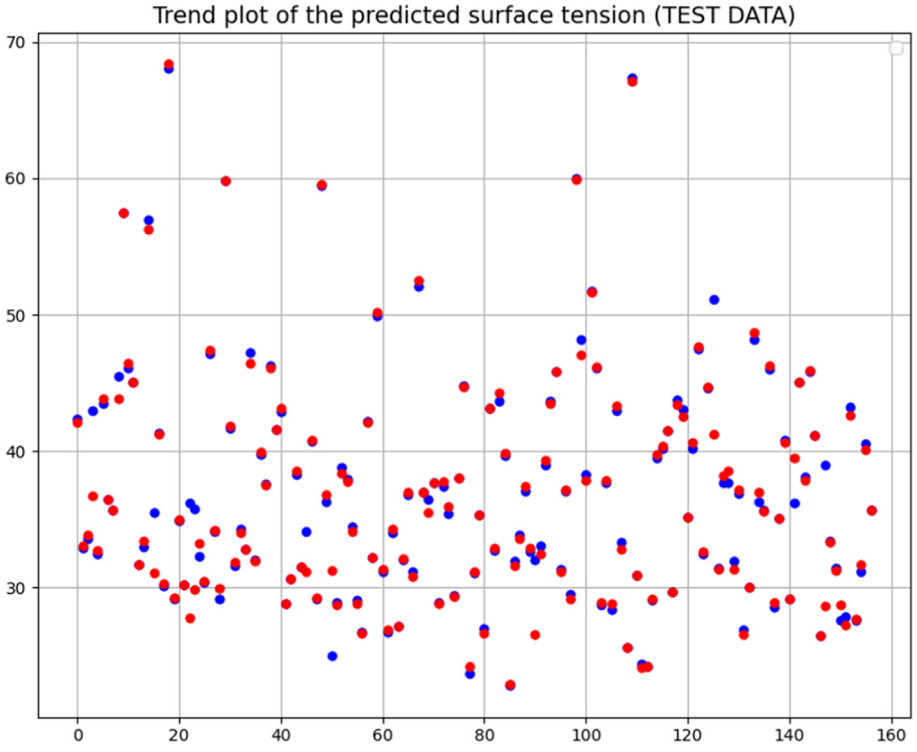

Fig. 5 Tendency of predicted surface tension – test data.

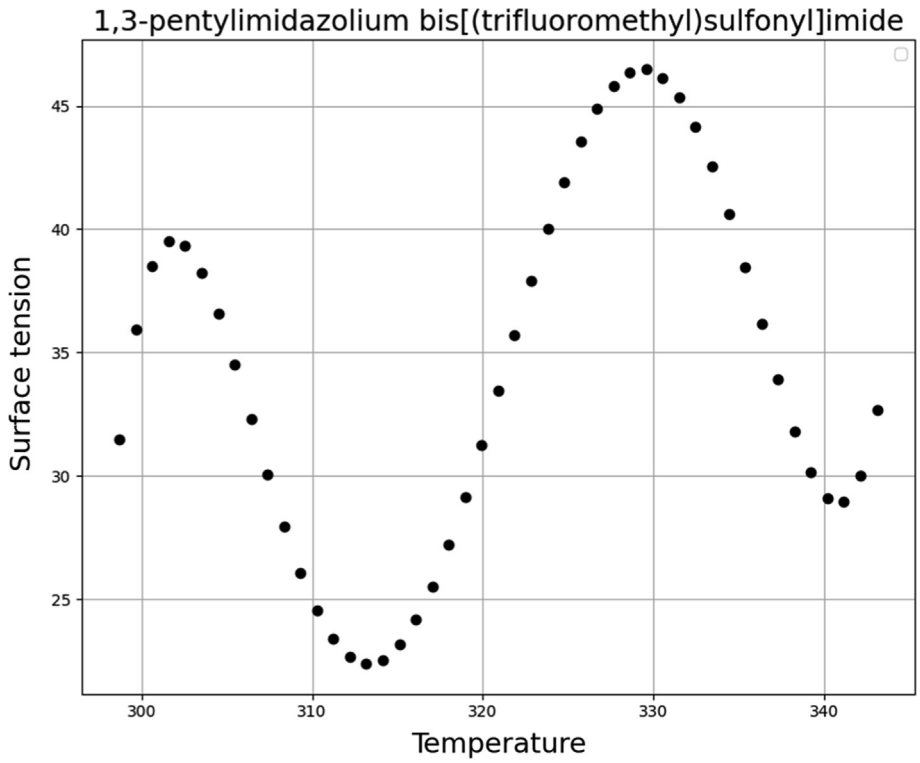

Fig. 6 Trends of Temperature (1,3-pentylimidazolium bis[(*trifluoromethyl*)*sulfonyl*]imide).

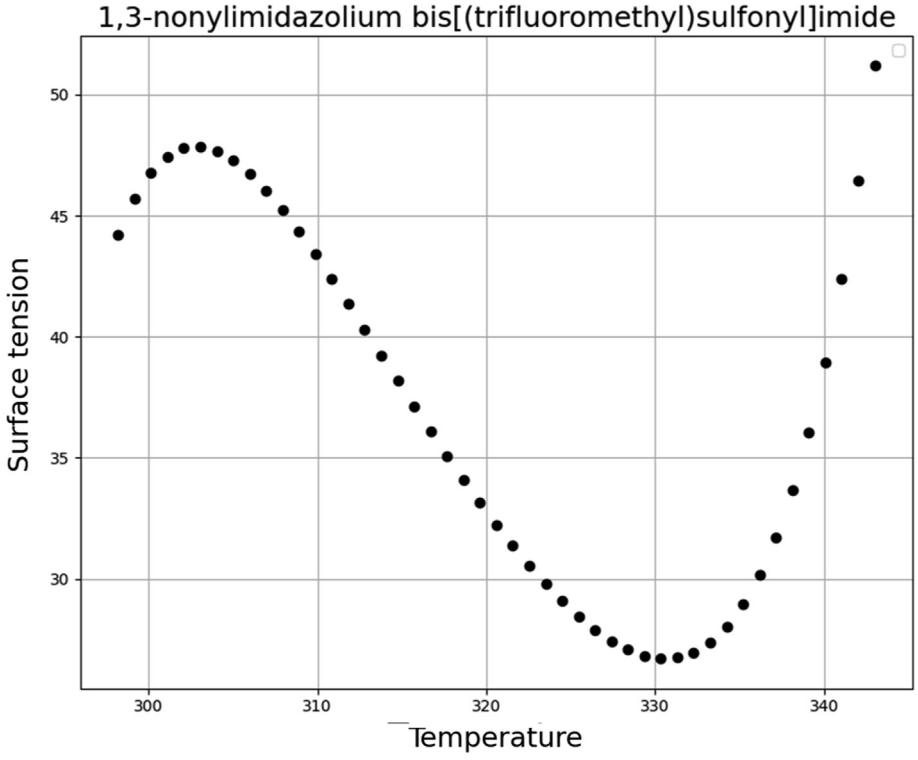

Fig. 7 Trends of Temperature (1,3-nonylimidazolium bis[(*trifluoromethyl*)*sulfonyl*]imide).

$$R^2 = 1 - \frac{\sum (y_i - x_i)^2}{\sum (x_i - \bar{x_i})^2}$$

MAPE as a common metrics, because of the independence.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} |\frac{x_i - y_i}{x_i}|$$

y_i and x_i are predicted and expected values.

Table 1 shows the final findings of the models. Figs. 1, 2, and 3 compare predicted and estimated values to illustrate the performance of AdaBoost-DT, AdaBoost-GPR and AdaBoost-SVR predictive models. Blue points in these graphs represent expected values in a training data set, while red points represent test data points. Comparison of outcomes corroborates that AdaBoost-GPR mathematical model demonstrates superior precision and generality owing to having higher R^2 and lower MAPE.

Investigation of the convergence between expected and predicted values is a useful way to evaluate the accuracy of model outcomes. Figs. 4 and 5 aims to comparatively illustrate the trend plot of the estimated surface tension via AdaBoost-GPR mathematical model in both modes (train and test data). As illustrated in the Figures, the presence of great convergence among the expected and estimated amounts of surface tension proves the validity of AdaBoost-GPR mathematical model for the mathematical estimation of the ILs' surface tension.

The influence of temperature as a significant operational parameter on the surface tension of 1,3-pentylimidazolium bis [(trifluoromethyl)sulfonyl] imide and 1,3-nonylimidazolium bis [(trifluoromethyl) sulfonyl] imide ILs are presented in Figs. 6 and 7. As expected, increment in temperature significantly reduces the intermolecular forces. Therefore, increase in temperature significantly facilitates the molecular movement of liquid, which results in declining the surface tension.

5. Conclusion

Investigations about the interfacial parameters of ILs (i.e., surface tension) are significantly increasing. Surface tension is a significant parameter of any liquid–gas interface, which its experimental measurement has been recently done by various researchers. Despite undoubted potential of experimental investigations in the measurement of surface tension, their effortful, expensive and prohibitive nature have motivated the scientists to precisely predict the ILs' properties via a dependable technique based on artificial intelligence. In this Study, we choose to use AdaBoost with various simple models including SVR, GPR, and DT. A genetic algorithm (GA) search is also used for feature selection and hyper-parameter tuning. The final R2-scores for boosted DT, GPR, and SVR are 0.849, 0.981, and 0.944, respectively. Accordingly, the ADABOOST-GPR is selected as the primary model for the study. MSE and MAPE error rates for this model are also 3.05 and 1.73E-02 that are better than two other models.

Acknowledgement

- The authors would like to acknowledge the Deanship of Scientific Research at Umm Al-Qura University, for supporting this work by Grant code: 22UQU4320545DSR25
- Authors would like to extend their appreciation to the deanship of scientific research at King Khalid University supporting this research through large group program under grant number RGP.2/64/43.

References

Amar, M.N., Zeraibi, N., Jahanbani Ghahfarokhi, A., 2020. Applying hybrid support vector regression and genetic algorithm to water alternating CO2 gas EOR. Greenhouse Gases Sci. Technol. 10, 613–630.

An, H., Landis, J.T., Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101. PJ Huber and EM Ronchetti. Robust Statistics. Wiley Series in Probability and Statistics. John Wiley & Sons, 2009..

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., 2013. API design for machine learning software: experiences from the scikit-learn project, arXiv preprint arXiv:1309.0238.

Dargahi-Zarandi, A., Hemmati-Sarapardeh, A., Shateri, M., Menad, N.A., Ahmadi, M., 2020. Modeling minimum miscibility pressure of pure/impure CO2-crude oil systems using adaptive boosting support vector regression: Application to gas injection processes. J. Petrol. Sci. Eng. 184, 106499.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer, pp. 1–15.

Esmaeili, H., Hashemipour, H., 2021. A simple correlation to predict surface tension of binary mixtures containing ionic liquids. J. Mol. Liq. 324, 114660.

Fehér, C., Kriván, E., Hancsók, J., Skoda-Földes, R., 2012. Oligomerisation of isobutene with silica supported ionic liquid catalysts. Green Chem. 14, 403–409.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Gardas, R.L., Coutinho, J.A., 2008. Applying a QSPR correlation to the prediction of surface tensions of ionic liquids. Fluid Phase Equilib. 265, 57–65.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Machine learning basics. Deep learning 1, 98–164.

Gouda, S.G., Hussein, Z., Luo, S., Yuan, Q., 2019. Model selection for accurate daily global solar radiation prediction in China. J. Cleaner Prod. 221, 132–144.

Hashemkhani, M., Soleimani, R., Fazeli, H., Lee, M., Bahadori, A., Tavalaeian, M., 2015. Prediction of the binary surface tension of mixtures containing ionic liquids using Support Vector Machine algorithms. J. Mol. Liq. 211, 534–552.

Járvás, G., Kontos, J., Babics, G., Dallos, A., 2018. A novel method for the surface tension estimation of ionic liquids based on COSMO-RS theory. Fluid Phase Equilib. 468, 9–17.

Jiang, Y., Jia, J., Li, Y., Kou, Y., Sun, S., 2021. Prediction of gas-liquid two-phase choke flow using Gaussian process regression. Flow Meas. Instrum. 81, 102044.

Kadavi, P.R., Lee, C.-W., Lee, S., 2018. Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing 10, 1252.

Keane, A., Forrester, A., Sobester, A., 2008. Engineering design via surrogate modelling: a practical guide. American Institute of Aeronautics and Astronautics, Inc..

Kecman, V., 2005. Support vector machines–an introduction. In: Support vector machines: theory and applications. Springer, pp. 1–47.

Kianfar, E., Mafi, S., 2021. Ionic liquids: properties, application, and synthesis. Fine Chem. Eng., 22–31

Lal, B., Qasim, A., Shariff, A.M., 2021. Ionic liquids in flow assurance. Springer.

Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18, 559–563.

Zia ul Mustafa, M., Bin Mukhtar, H., Md Nordin, N.A.H., Mannan, H.A., Nasir, R., Fazil, N., 2019. Recent developments and applications of ionic liquids in gas separation membranes. Chem. Eng. Technol. 42, 2580–2593.

Mahandra, H., Faraji, F., Ghahreman, A., 2021. Novel extraction process for gold recovery from thiosulfate solution using phosphonium ionic liquids. ACS Sustainable Chem. Eng. 9, 8179–8185.

Mathuria, M., 2013. Decision tree analysis on j48 algorithm for data mining, Intrenational Journal of Advanced Research in Computer Science and Software. Engineering 3.

Mirkhani, S.A., Gharagheizi, F., Farahani, N., Tumba, K., 2013. Prediction of surface tension of ionic liquids by molecular approach. J. Mol. Liq. 179, 78–87.

Mjalli, F.S., Vakili-Nezhaad, G., Shahbaz, K., AlNashef, I.M., 2014. Application of the Eötvos and Guggenheim empirical rules for predicting the density and surface tension of ionic liquids analogues. Thermochim Acta 575, 40–44.

Moosaei, H., Ketabchi, S., Razzaghi, M., Tanveer, M., 2021. Generalized twin support vector machines. Neural Process. Lett. 53, 1545–1564.

Mousavi, S.-P., Atashrouz, S., Amar, M.N., Hadavimoghaddam, F., Mohammadi, M.-R., Hemmati-Sarapardeh, A., Mohaddespour, A., 2021. Modeling surface tension of ionic liquids by chemical structure-intelligence based models. J. Mol. Liq. 342, 116961.

Quinonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression, The. Journal of Machine Learning Research 6, 1939–1959.

Sakar, A., Mammone, R.J., 1993. Growing and pruning neural tree networks. IEEE Trans. Comput. 42, 291–299.

Shang, D., Liu, X., Bai, L., Zeng, S., Xu, Q., Gao, H., Zhang, X., 2017. Ionic liquids in gas separation processing, Current Opinion in Green and Sustainable. Chemistry 5, 74–81.

Shi, J.Q., Choi, T., 2011. Gaussian process regression analysis for functional data. CRC Press.

Shojaeian, A., 2018. New experimental and modeling based on the N-Wilson-NRF equation for surface tension of aqueous alkanolamine binary mixtures. J. Mol. Liq. 254, 26–33.

Shojaeian, A., Asadizadeh, M., 2020. Prediction of surface tension of the binary mixtures containing ionic liquid using heuristic approaches; an input parameters investigation. J. Mol. Liq. 298, 111976.

Song, Y.-Y., Ying, L., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry 27, 130.

Vapnik, V., 1999. The nature of statistical learning theory. Springer science & business media.

Wang, X., Chi, Y., Mu, T., 2014. A review on the transport properties of ionic liquids. J. Mol. Liq. 193, 262–266.

Wang, L., Zheng, C., Zhou, W., Zhou, W.-X., 2020. A new principle for tuning-free Huber regression. Statistica Sinica.

Wilson, A.G., Knowles, D.A., Ghahramani, Z., 2011. Gaussian process regression networks, arXiv preprint arXiv:1110.4411.

Wu, C., Khan, Z., Ioannidis, S., Dy, J.G., 2020. Deep Kernel Learning for Clustering. In: Proceedings of the 2020 SIAM International Conference on Data Mining, pp. 640–648.

Xu, M., Watanachaturaporn, P., Varshney, P.K., Arora, M.K., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sens. Environ. 97, 322–336.

Zhou, Z.-H., 2019. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.