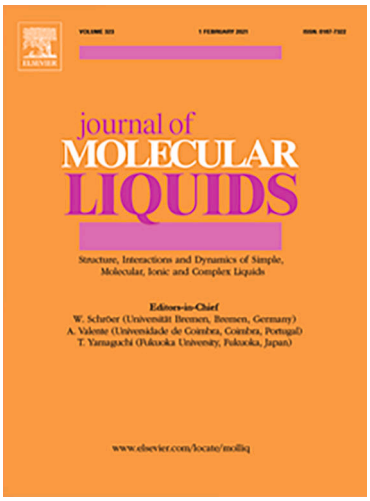
Journal Pre-proofs



Optimization and validation of drug solubility by development of advanced artificial intelligence models

Yaoyang Liu, Drai Ahmed Smait, Abbas Yaseen Naser, Farag M. A. Altalbawy, Hala Bahri, Ali Abdul Kadhim Ruhaima, Thura Zayad Fathallah, Salema K. Hadrawi, Refad E. Alsaddon, Abdullah Alshetaili, Amal M. Alsubaiyel

PII: S0167-7322(22)02652-6
DOI: https://doi.org/10.1016/j.molliq.2022.121113
Reference: MOLLIQ 121113

To appear in: *Journal of Molecular Liquids*

Received Date: 29 November 2022
Revised Date: 13 December 2022
Accepted Date: 18 December 2022

Please cite this article as: Y. Liu, D. Ahmed Smait, A. Yaseen Naser, F. M. A. Altalbawy, H. Bahri, A. Abdul Kadhim Ruhaima, T. Zayad Fathallah, S.K. Hadrawi, R.E. Alsaddon, A. Alshetaili, A.M. Alsubaiyel, Optimization and validation of drug solubility by development of advanced artificial intelligence models, *Journal of Molecular Liquids* (2022), doi: https://doi.org/10.1016/j.molliq.2022.121113

Optimization and validation of drug solubility by development of advanced artificial intelligence models

Yaoyang Liu [1,2,*], Drai Ahmed Smait[3], Abbas Yaseen Naser[4], Farag M. A. Altalbawy[5,6], Hala Bahri[7], Ali Abdul Kadhim Ruhaima[8], Thura Zayad Fathallah[9], Salema K. Hadrawi[10], Refad E. Alsaddon [11], Abdullah Alshetaili [12], Amal M. Alsubaiyel[13,*]

[1] Jiangxi Cas pharmaceutical Engineering Technology Co.,LTD, Nanchang, Jiangxi, 330000, China

[2] Chinese Academy of Sciences Hatch Innovation Investment Co., Ltd, Beijing, 100000, China

[3] The University of Mashreq, Baghdad, Iraq

[4] Information technology Uint , Al mustaqbal University college , Babylon 51001, iraq

[5] National Institute of Laser Enhanced Sciences (NILES), University of Cairo, Giza 12613, Egypt

[6]Department of Chemistry, University College of Duba, Univeristy of Tabuk, Tabuk, Saudi Arabia

[7] College of pharmacy, Al-Farahidi university, Iraq

[8] AL-Nisour University College, Baghdad, Iraq

[9] Department of Pharmacy, AlNoor University College, Bartella, Iraq

[10] Refrigeration and Air-conditioning Technical Engineering Department, College of Technical Engineering, The Islamic University, Najaf, Iraq

[11] Department of Prosthetic Dental Technology, Hilla University college, Babylon, Iraq

[12] Department of Pharmaceutics, College of Pharmacy, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

[13]Department of Pharmaceutics, College of Pharmacy, Qassim University, Buraidah 52571, Saudi Arabia.

*Corresponding author: liu.yaoyang@yahoo.com, alsubaiyela@gmail.com

Abstract

Over the last ten years, the application of novel mathematical models of Machine Learning employed to model the solubility of drugs especially anticancer drugs, in supercritical carbon dioxide ($ScCO_2$) system has gained remarkable popularity. In this research, three distinct ensemble models have been employed on the data as a novel method for busulfan as anticancer drug for the first time, based on decision trees, including Random Forest (RF), Gradient Boosting Trees (GBRT), and Extremely Randomized Tree (ERT) to predict the solubility of busulfan as an

anticancer drug. The dataset has two input parameters, T=Temperature and P=Pressure, and Y= Solubility is the single output. After implementing and tuning these ensemble models' hyper parameters, the performance has been assessed through several metrics. All three models show R-squared score of more than 0.9, but in terms of RMSE, the error rates are 1.80E-04, 1.72E-04, and 1.03E-04 for RF, ERT, and GBRT models, respectively. Also, MAPE metrics 4.51E-01, 4.87E-01, and 3.62E-01 errors had found for RF, ERT, and GBRT models, respectively. GBRT has been selected as the best model due to the less rate of RMSE and MAPE. An analysis has also been performed to find the optimal amount of solubility, which can be considered the (x1=38.3, x2=333.1, Y=1.36E-03) vector.

Keywords: Anticancer drug; Supercritical fluids; Simulation; Machine learning; Model validation

1 INTRODUCTION

True perception of cancer biology is an important factor towards developing efficacious anticancer therapeutic agent to treat disparate malignant tumors [1, 2]. Apart from acceptable efficiency, manageable toxicity profile of anticancer drugs is an indisputable factor in the new era of oncology drug development due to its significant effect on the quality of life of patients [3, 4]. Therefore, efficient optimization of drug dosage is an important challenge in anticancer drug development.

Supercritical carbon dioxide ($ScCO_2$) has been recently of extensive utilization in biochemical/pharmaceutical industries as a promising solvent because of its remarkable characteristics like near-ambient critical temperatures, safety of operation, inflammability and versatility [5-7]. Drug solubility and materials behavior are two important parameters, which can be achieved by $ScCO_2$ technology for the manufacturing of nano/microparticles of anticancer drugs [8-10] . The prosperous use of $ScCO_2$ relies on the true perception of solute solubility. In recent years, numerous theoretical/experimental investigations have been conducted to predict the solubility of disparate materials in supercritical fluids (SCFs) [11].

Empirical correlations (ECs) are known as one of the promising thermodynamic techniques for predicting the amounts of drug solubility. ECs models have been of paramount attention owing to their simplicity of implementation for the solubility data. These predictive models often possess some fitting parameters, which are able to be specified by means of numerical techniques. Despite noteworthy advantages, ECs do not have sufficient accuracy to be applied for an extensive range of drugs [12-14]. With the goal of enhancing the precision of drug solubility prediction in $ScCO_2$, software-based computational procedures such as machine learning (ML) can be employed. ML-based predictive models, which are based on artificial intelligence (AI), profoundly depend on the

training data. The solubility prediction of drugs in these models take place applying the trained machine [15-19].

There are numerous problems in various scientific fields where there is no clear or linear relationship between input and output values, or the number of available laboratory data is too small or too large to be easily discovered with a human factor. In such situations, machine learning (ML), as the strongest knowledge available in this field, can help to explore these relationships with different tools [20-23].

The decision tree a widely used ML tools. The decision tree (DT) is a structure that generates trees for regression or classification problems. As it generates the associated decision tree, the dataset is gradually segmented into smaller subsets. The output is a tree with both leaf and decision nodes. The decision node may have two or more branches, with each branch representing a possible outcome for the criterion in question. The Leaf node stands [24-26].

Also, Ensemble learning [27] is another popular mixed model that excels in situations where modeling is too difficult to be accomplished with a single approach. They use a variety of learning algorithms to collectively arrive at a more precise estimate. The goal of ensemble models is to decrease bias and variance by combining multiple machine learning or statistical algorithms into a single model [28, 29]. The aim of this research is to model the solubility of busulfan as anticancer drug through machine learning using different models. Boosting and bagging are two types of ensembles learning methods. Boosting is an interdependent base estimator seeking sub-sequential learning algorithm. There will almost certainly be less bias in the overall performance compared to using a single base estimator. Bagging is an approach to parallel learning that averages multiple base estimators in an effort to reduce variance [30-32].

We used three ensemble estimators in this work, all based on DT. Random Forest and Extremely randomized tress are bagging and gradient boosting trees are boosting methods selected for this research.

2 DATA SET

The experimental dataset of this study has two input features P and T and Y= Solubility is the single output that is shown in Table 1. Figure 1 demonstrates the distribution of columns of data set [15].

Table 1- the whole dataset

No.	X1=P	X2=T	y
1	12	308	6.75E-05
2	16	308	1.32E-04
3	20	308	1.77E-04
4	24	308	2.23E-04
5	28	308	2.77E-04
6	32	308	3.41E-04
7	36	308	3.93E-04
8	40	308	4.29E-04
9	12	318	3.46E-05
10	16	318	1.13E-04
11	20	318	1.94E-04
12	24	318	2.70E-04
13	28	318	3.85E-04
14	32	318	5.01E-04
15	36	318	5.67E-04
16	40	318	7.07E-04
17	12	328	2.20E-05
18	16	328	7.65E-05
19	20	328	2.23E-04
20	24	328	3.09E-04
21	28	328	4.88E-04
22	32	328	6.60E-04
23	36	328	8.16E-04
24	40	328	8.81E-04
25	12	338	1.35E-05

26	16	338	6.33E-05
27	20	338	2.07E-04
28	24	338	3.73E-04
29	28	338	6.07E-04
30	32	338	8.20E-04
31	36	338	9.01E-04
32	40	338	1.28E-03

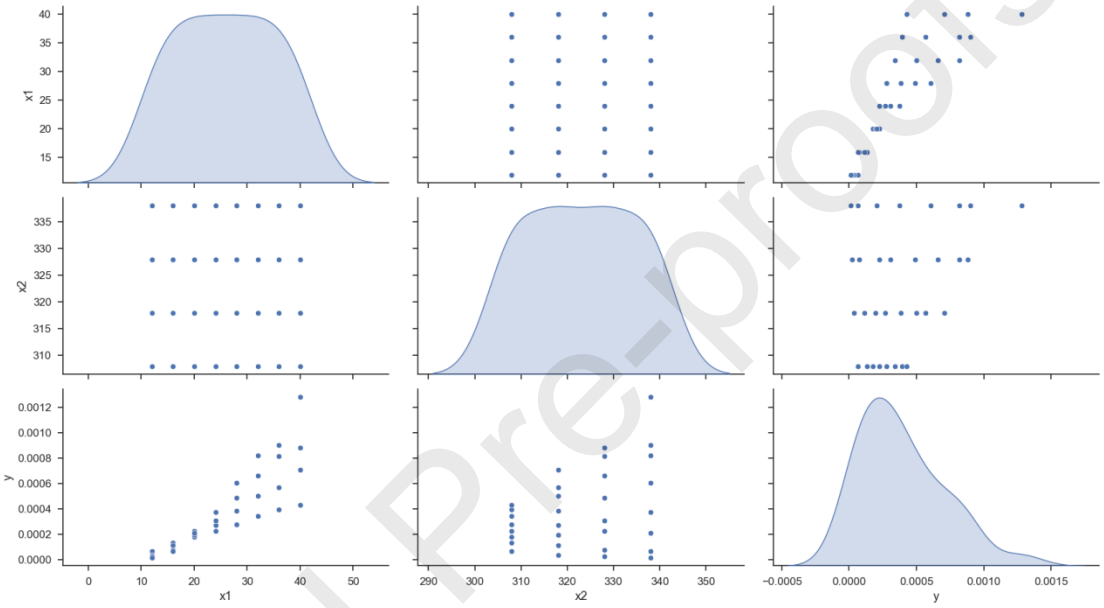


Figure 1- Data Distribution

3 METHODOLOGY

3.1 RANDOM FOREST (RF)

The Random Forest method introduced by Breiman is a form of ensemble methods [33]. In terms of general-purpose learning algorithms, RF is one of the most reliable and efficient. The Random Forest algorithm has been shown to deliver excellent results while utilizing minimal computing power. RF outperforms single-tree algorithms like CART in terms of speed. Compared to more time-consuming and resource-intensive algorithms, this one performs as well as it does [34, 35].

In this regression model, decision tree estimators $h(x,\theta_k)$, k ∈ {1,…,K} take on numerical values according to the random vectors $\{\theta_k\}$. Each piece of data in the training set is randomly selected from a joint distribution of the random vectors *(X,Y)*, where *X* denotes what was observed and *Y* denotes what was expected. The Classification and Regression Trees (CART) [36] algorithm is used to grow individual trees. Listed below is the random forest regression algorithm [37, 38]:

- For *i in {1,...,B}* :
    - Generate a bootstrapped subset Z^* of including *N* data points from training subset.
    - Create a decision T_i with Z^* by iteratively repeat the steps below for every leaf node as late as the node size n_{min} is gained.
        - Choose *m* input variables randomly from *p* input variables.
        - Choose the optimal split point amongst the *m*.
        - Divide the node into multiple Childs.
- Output: the collection of Decision trees $\{T_i\}_1^B$ (Ensemble)
- To estimate an unseen input such as *x*:

$$\bar{f}_{rf}^{B} = \frac{1}{B}\sum_{i=1}^{B} T_i(x)$$

3.2 EXTREMELY RANDOMIZED TREES (ERT)

Extremely randomized trees (ERT) is another Decision Tree based model that develops an ensemble of Decision Trees with no pruning in the traditional top-to-bottom tree growth fashion [39]. ERT, alike other tree-based ensemble models, constructs a group of DTs, but with a heavy focus on randomization to reduce variance without increasing bias too much [40].

The tree growing mechanism introduces randomness by creating split nodes with features and cut-points selected randomly. In this case, the entire learning set is used to build the tree, rather than a bootstrap sample. As a result of the randomization and average of ensembles, DTs' latent variance should be reduced while bias should be avoided by using the original learning sample instead of replicas from bootstraps. Three parameters are important in this regard: the minimum samples for a node-split (n_{min}), the count of randomly selected input features at each node (K). Once the predetermined number of ensemble trees has been generated, a majority vote is used to aggregate their respective predictions, resulting in the final estimated values for the ERT model [41, 42].

3.3 GRADIENT BOOSTING TREES (GBRT)

In the same way that new base estimators learn from older ones, GBRT is another ensemble method. As more weak estimators are added to the model, the error rate decreases. Meanwhile, the GBRT algorithm sets a ratio for base estimators called the learning rate to prevent overfitting. The final prediction from the GBRT model is the sum of the predictions from the weak estimators times the learning rate [43]. The steps involved in the GBRT method are shown in the following algorithm:

Set initial value $F_0(x) = argmin_p \sum_{i-1}^{N} L(y_i, P)$

For $t\ in\ \{1,2,\ldots,M\}$:

1. Determine the negative gradient

$$\overline{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F_{x_i}}\right]$$

2. Create a tree model

$$a_t = argmin_{a,\beta} \sum_{i=1}^{N} [\overline{y} - \beta h(x_i, a_t]^2$$

3. Pick a step size for the gradient descent as

$$p_t = argmin_p \sum_{i=1}^{N} L(y_i, Ft-1(x_i) + ph(x_i, a))$$

4. Update the estimation of *F(x)*

$$F_t(x) = F_t - 1(x) + p_t h(x, a_t)$$

Output: $F_t(x)$

4 RESULTS

Standard evaluation criteria were used to fine-tune the hyper-parameters of the three presented models. This section details these requirements.

R^2 (Coefficient of Determination) is a common metric for evaluating the performance of prediction results. It indicates how closely the regression analyses match the observed data's trends [44].

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

The RMSE is assessed using the sample standard deviation of the separation between the model generated and the actual values. It gives us a way to evaluate how well our predictions are doing [45].

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

Because of its interpretability and scale independence, MAPE is one of the popular metrics. MAPE is a statistical metric for determining how accurate a predictive model is. MAPE discusses the error size as float in the interval of (0 , 1) [46].

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

y_i and x_i are model generated and experimental values, respectively. Also, x_i denotes the mean of actual data. n denotes the count of datapoints in the dataset.

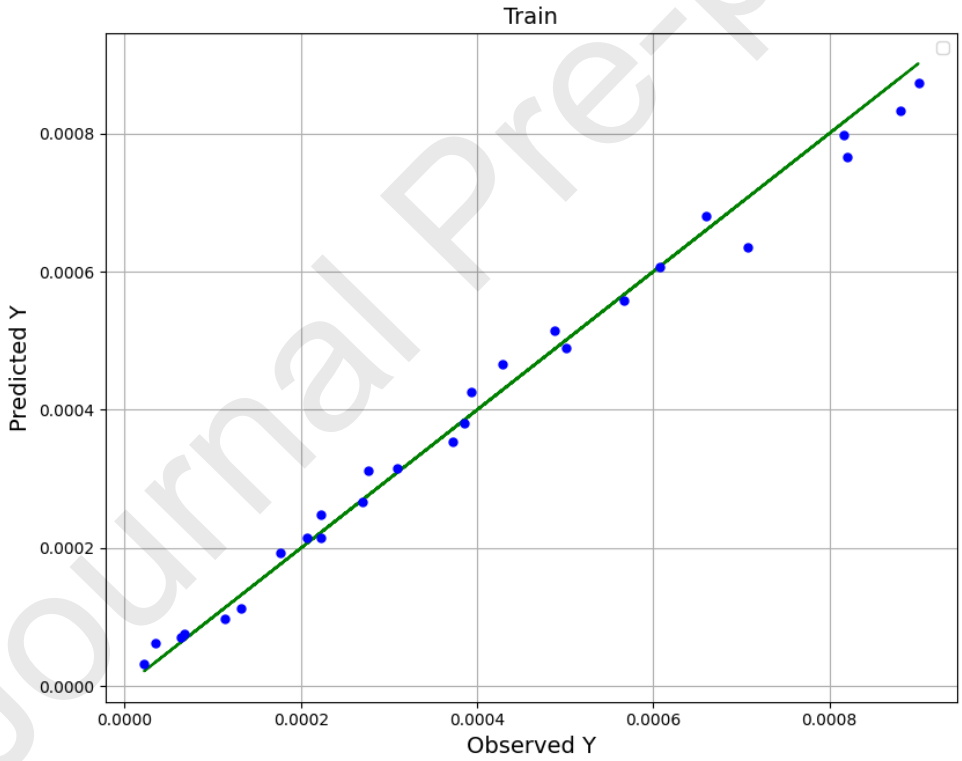


Figure 2- RF training phase

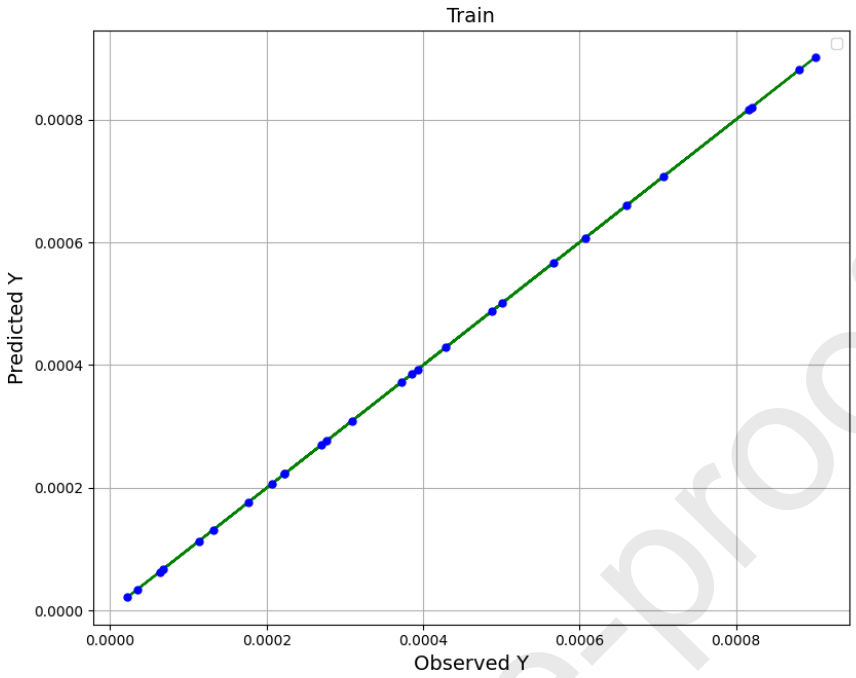


Figure 3- ERT training phase

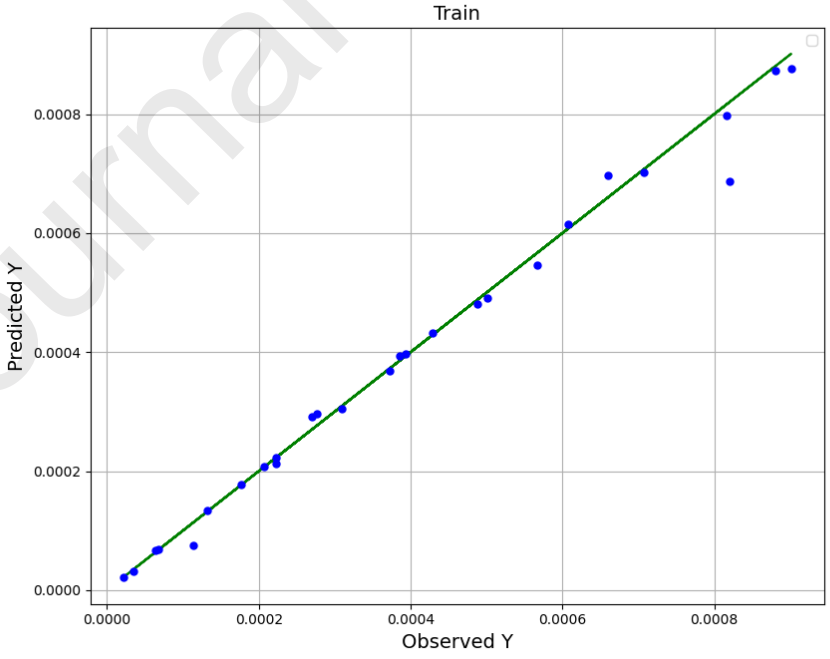


Figure 4- GBRT training phase

In Figure 5, we can see a three-dimensional representation of the inputs being projected onto a single channel of output. Optimal values of X1, X2 confirm this, as they are close to their upper bounds, showing that increasing the value of both features will roughly increase output.

The statistical analysis results of the fitting for three developed models (RF, ERT and GBRT) to prognosticate the busulfan solubility is displayed in Table 2. Evidently, all models have acceptable precision to predict the experimental data. Compared with other predictive models, GBRT model shows greater model validation due to its high R-Squared and RMSE values (0.987 and 1.03E-04).

Table 2- The Performance of final optimized models

Models	R-Squared	RMSE	MAPE
RF	0.988	1.80E-04	4.51E-01
ERT	0.945	1.72E-04	4.87E-01
GBRT	0.987	1.03E-04	3.62E-01

Figures 5, 6 and 7 respectively demonstrate the three-dimensional plot for evaluating the simultaneous effect of X1 (pressure) and X2 (temperature) on the Y (solubility) and the two-dimensional evaluation of temperature and pressure on the busulfan solubility, respectively. Analysis of the figures corroborates that despite the positive impact of pressure on the solubility amounts of busulfan for all evaluated isotherms, its influence improves by increasing the temperature. This reason can be justified due to this fact that the increment of temperature eventuates in a significant reduction in the molecular distance between the CO_2 molecules and modifying the system towards the liquid state. Shifting the system towards the liquid state automatically increases the density of the system, which ultimately causes an enhancement in the solubilizing power of CO_2. Therefore, increment in the solubilizing power of CO_2 eventuates in

improving the dissolution of busulfan in $ScCO_2$. Additionally, stricter evaluation of the predicted busulfan solubility data for all the presented isotherms demonstrate an abnormal manner at the temperature around 20 MPa where the impact of temperature became reversed. At the pressure lower than this point, increasing the temperature reduces the busulfan solubility because of density reduction, while for the pressures more than 20 MPa, the influence of temperature returns to an enhancing trend. Therefore, an important parameter to determine the value of busulfan solubility in $ScCO_2$ is cross-over pressure at the pressure of almost 20 MPa where the temperature influence on the busulfan solubility reverses. It can be concluded form the outcomes that temperature has a paradoxical impact on the busulfan solubility in $ScCO_2$ system. At the pressures lower than 20 MPa, increasing the temperature decreases the busulfan solubility. At these pressures, increase in the temperature reduces the density parameter despite increasing the sublimation pressure. At these pressures, the effect of density decrement is more than increasing the sublimation pressure and thus, increase in the temperature declines the busulfan solubility. At the pressures higher than cross-over pressure (20MPa), the impact of pressure sublimation overcomes the effect of density. Hence, increase in the temperature at these pressures increases the solubility value of busulfan in the $ScCO_2$ system [15]. The optimized values for pressure and temperature to obtain the highest solubility of busulfan is presented in Table 3.

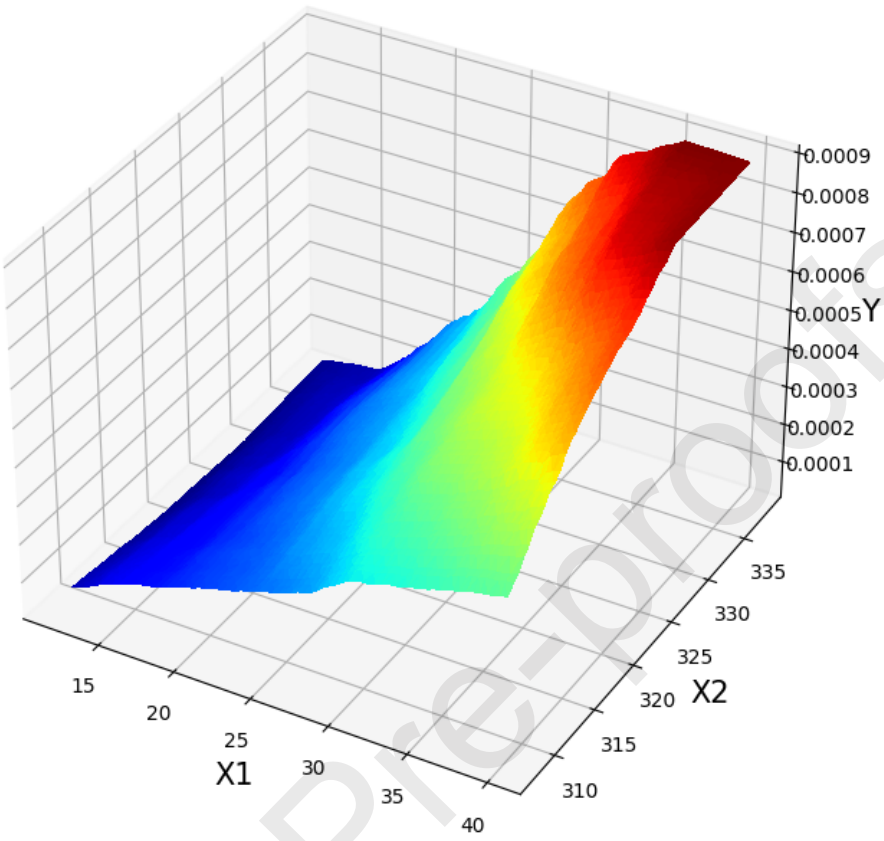


Figure 5- the 3D inputs/outputs projection (GBRT Model)

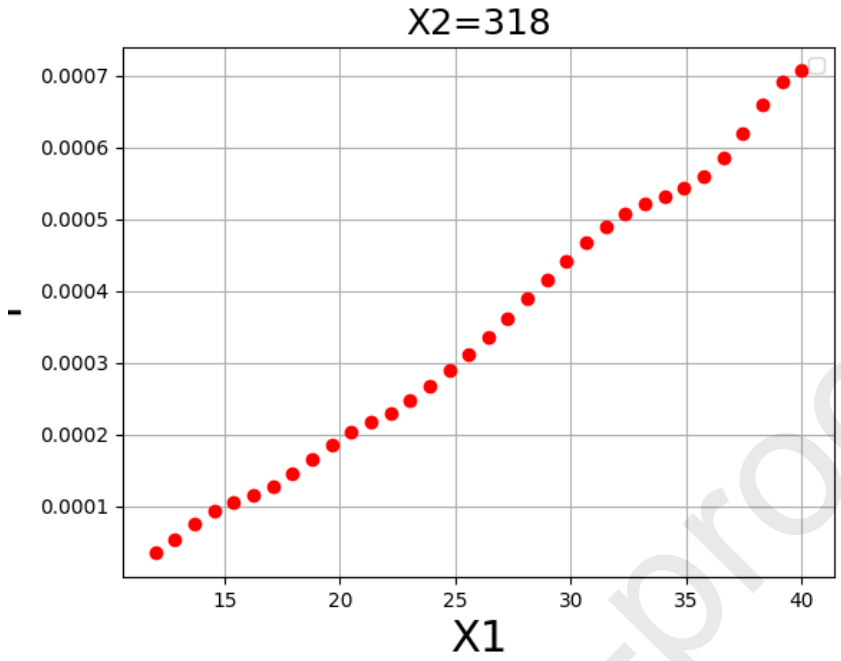


Figure 6- The 2D trends for X1

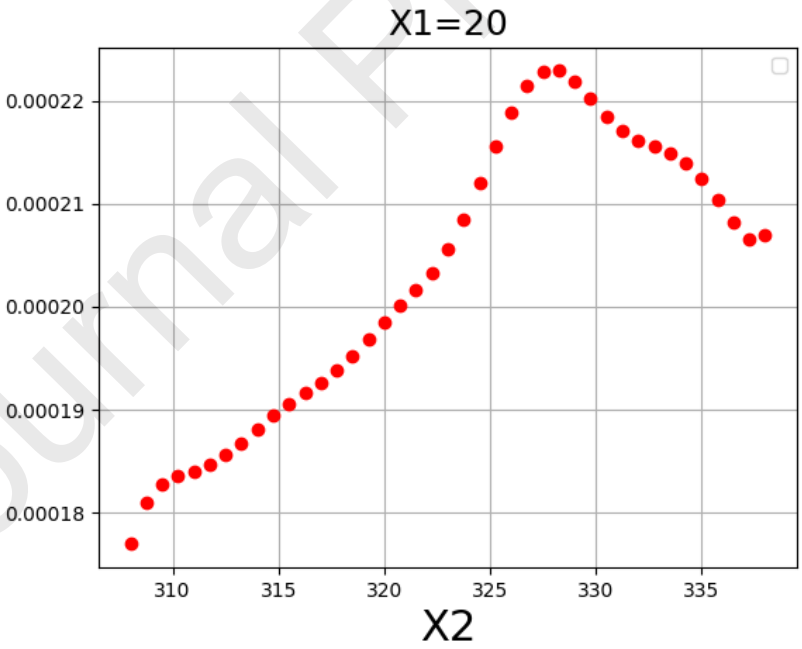


Figure 7- The 2D trends for X2

Table 3- final optimal values of the parameters for maximum solubility

X1=P	X2=T	Y
38.3	333.1	1.36E-03

5 CONCLUSION

In this paper, the optimum value of busulfan solubility in supercritical carbon dioxide ($ScCO_2$) system was aimed to be predicted. For this purpose, three predictive novel models for busulfan based on artificial intelligence and machine learning approach were developed and trained applying the experimental data. This study used different decision tree-based ensemble estimators: RF, ERT, and GBRT. Following the implementation and tuning of these ensemble models' hyper parameters, their overall performance have been evaluated through a variety of criteria. Although all three models show an R-squared score more than 0.9, the RMSE error rates for the RF, ERT, and GBRT models are 1.80E-04, 1.72E-04, and 1.03E-04, respectively, for the RF, ERT, and GBRT models. In addition, MAPE metrics of 4.51E-01, 4.87E-01, and 3.62E-01 errors were obtained for the RF, ERT, and GBRT estimators, respectively, according to the results. GBRT has been selected as the best model due to the less rate of RMSE and MAPE. The optimal solubility amount, expressed by the vector (x1=38.3, x2=333.1, Y=1.36E-03), has also been analyzed.

• Data Availability Statement:

All data are available within the published paper.

• Supplementary information:

No supplementary

• **Ethical approval:**

Not applicable

Author contribution:

Yaoyang Liu: conception, experimental design, carrying out measurements and manuscript composition, writing

Drai Ahmed Smait: writing, investigation, experimental design, carrying out measurements, modelling and analysis

Abbas Yaseen Naser: resources, experimental design, carrying out measurements, writing, analysis

Farag M. A. Altalbawy: investigtaion, experimental design, carrying out measurements and writing, formal analysis

Hala Bahri: writing and editing, investigation, validation

Ali Abdul Kadhim Ruhaima: writing and editing, investigation, validation, resources

Thura Zayad Fathallah: writing and editing, investigation, validation, resources, analysis

Salema K. Hadrawi: writing and editing, investigation, resources, analysis

Refad E. Alsaddon: writing and editing, investigation, resources, analysis

Abdullah Alshetaili: writing and editing, investigation, resources, analysis

Amal M. Alsubaiyel: writing and editing, investigation, validation, resources, analysis, supervision

6 REFERENCES

1. Sharma, P., et al., *Innovation in cancer therapeutics and regulatory perspectives.* Medical Oncology, 2022. **39**(5): p. 1-12.
2. Sarikaya, I., *Biology of cancer and PET imaging: pictorial review.* Journal of Nuclear Medicine Technology, 2022. **50**(2): p. 81-89.
3. Gangadhar, T.C. and R.H. Vonderheide, *Mitigating the toxic effects of anticancer immunotherapy.* Nature reviews Clinical oncology, 2014. **11**(2): p. 91-99.
4. Basak, D., et al., *Comparison of Anticancer Drug Toxicities: Paradigm Shift in Adverse Effect Profile.* Life, 2021. **12**(1): p. 48.
5. Duarte, A.R.C., et al., *Solubility of flurbiprofen in supercritical carbon dioxide.* Journal of Chemical & Engineering Data, 2004. **49**(3): p. 449-452.
6. Subramaniam, B., R.A. Rajewski, and K. Snavely, *Pharmaceutical processing with supercritical carbon dioxide.* Journal of pharmaceutical sciences, 1997. **86**(8): p. 885-890.
7. Guo, J.-Q., et al., *A systematic review of supercritical carbon dioxide (S-CO2) power cycle for energy industries: Technologies, key issues, and potential prospects.* Energy Conversion and Management, 2022: p. 115437.
8. Savjani, K.T., A.K. Gajjar, and J.K. Savjani, *Drug solubility: importance and enhancement techniques.* ISRN pharmaceutics, 2012. **2012**: p. 195727-195727.
9. Göke, K., et al., *Novel strategies for the formulation and processing of poorly water-soluble drugs.* European Journal of Pharmaceutics and Biopharmaceutics, 2018. **126**: p. 40-56.
10. Wang, S.-W., S.-Y. Chang, and C.-M. Hsieh, *Measurement and modeling of solubility of gliclazide (hypoglycemic drug) and captopril (antihypertension drug) in supercritical carbon dioxide.* The Journal of Supercritical Fluids, 2021. **174**: p. 105244.
11. Nguyen, H.C., et al., *Computational prediction of drug solubility in supercritical carbon dioxide: Thermodynamic and artificial intelligence modeling.* Journal of Molecular Liquids, 2022. **354**: p. 118888.
12. May, P.M. and D. Rowland, *Thermodynamic modeling of aqueous electrolyte systems: current status.* Journal of Chemical & Engineering Data, 2017. **62**(9): p. 2481-2495.
13. Ushiki, I., R. Fujimitsu, and S. Takishima, *Predicting the solubilities of metal acetylacetonates in supercritical CO2: Thermodynamic approach using PC-SAFT.* The Journal of Supercritical Fluids, 2020. **164**: p. 104909.
14. Mota, F.L., et al., *Temperature and solvent effects in the solubility of some pharmaceutical compounds: measurements and modeling.* European Journal of Pharmaceutical Sciences, 2009. **37**(3-4): p. 499-507.
15. Zhu, H., et al., *Machine learning based simulation of an anti-cancer drug (busulfan) solubility in supercritical carbon dioxide: ANFIS model and experimental validation.* Journal of Molecular Liquids, 2021. **338**: p. 116731.
16. Öztürk, A.A., A.B. Gündüz, and O. Ozisik, *Supervised machine learning algorithms for evaluation of solid lipid nanoparticles and particle size.* Combinatorial Chemistry & High Throughput Screening, 2018. **21**(9): p. 693-699.
17. Ding, Y., et al., *Artificial intelligence based simulation of Cd (II) adsorption separation from aqueous media using a nanocomposite structure.* Journal of Molecular Liquids, 2021. **344**: p. 117772.
18. Staszak, M., *Artificial intelligence in the modeling of chemical reactions kinetics.* Physical Sciences Reviews, 2020.
19. Tai, X.Y., et al., *The future of sustainable chemistry and process: Convergence of artificial intelligence, data and hardware.* Energy and AI, 2020. **2**: p. 100036.

20. Alpaydin, E., *Introduction to machine learning*. 2020: MIT press.
21. Bishop, C.M. and N.M. Nasrabadi, *Pattern recognition and machine learning*. Vol. 4. 2006: Springer.
22. El Naqa, I. and M.J. Murphy, *What is machine learning?*, in *machine learning in radiation oncology*. 2015, Springer. p. 3-11.
23. Schapire, R.E., *The boosting approach to machine learning: An overview.* Nonlinear estimation and classification, 2003: p. 149-171.
24. Mathuria, M., *Decision tree analysis on j48 algorithm for data mining.* Intrenational Journal of Advanced Research in Computer Science and Software Engineering, 2013. **3**(6).
25. Quinlan, J.R., *Learning decision tree classifiers.* ACM Computing Surveys (CSUR), 1996. **28**(1): p. 71-72.
26. Safavian, S.R. and D. Landgrebe, *A survey of decision tree classifier methodology.* IEEE transactions on systems, man, and cybernetics, 1991. **21**(3): p. 660-674.
27. Polikar, R., *Ensemble based systems in decision making.* IEEE Circuits and systems magazine, 2006. **6**(3): p. 21-45.
28. Ribeiro, M.H.D.M. and L. dos Santos Coelho, *Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series.* Applied Soft Computing, 2020. **86**: p. 105837.
29. Seyghaly, R., et al. *Interference Recognition for Fog Enabled IoT Architecture using a Novel Tree-based Method*. in *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. 2022. IEEE Computer Society.
30. Borra, S. and A. Di Ciaccio, *Improving nonparametric regression methods by bagging and boosting.* Computational Statistics & Data Analysis, 2002. **38**(4): p. 407-420.
31. Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.
32. Maclin, R. and D. Opitz, *An empirical evaluation of bagging and boosting.* AAAI/IAAI, 1997. **1997**: p. 546-551.
33. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
34. Ahmad, M.W., M. Mourshed, and Y. Rezgui, *Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption.* Energy and Buildings, 2017. **147**: p. 77-89.
35. Peters, J., et al., *Random forests as a tool for ecohydrological distribution modelling.* ecological modelling, 2007. **207**(2-4): p. 304-318.
36. Breiman, L., et al., *Classification and regression trees*. 2017: Routledge.
37. Trevor, H., T. Robert, and F. Jerome, *The elements of statistical learning: data mining, inference, and prediction*. 2009, Spinger.
38. Verikas, A., A. Gelzinis, and M. Bacauskiene, *Mining data with random forests: A survey and results of new tests.* Pattern recognition, 2011. **44**(2): p. 330-349.
39. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.
40. Song, J., et al., *Potential of ensemble learning to improve tree-based classifiers for landslide susceptibility mapping.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020. **13**: p. 4642-4662.
41. Wehenkel, L., D. Ernst, and P. Geurts, *Ensembles of extremely randomized trees and some generic applications.* Proceedings of robust methods for power system state estimation and load forecasting, 2006.
42. Kocev, D. and M. Ceci. *Ensembles of extremely randomized trees for multi-target regression*. in *International Conference on Discovery Science*. 2015. Springer.

43. Elith, J., J.R. Leathwick, and T. Hastie, *A working guide to boosted regression trees.* Journal of animal ecology, 2008. **77**(4): p. 802-813.
44. Gouda, S.G., et al., *Model selection for accurate daily global solar radiation prediction in China.* Journal of cleaner production, 2019. **221**: p. 132-144.
45. Zang, H., et al., *Application of functional deep belief network for estimating daily global solar radiation: A case study in China.* Energy, 2020. **191**: p. 116502.
46. Kim, S. and H. Kim, *A new metric of absolute percentage error for intermittent demand forecasts.* International Journal of Forecasting, 2016. **32**(3): p. 669-679.

Highlight:

1- Solubility enhancement through advanced artificial intelligence model
2- Applying different modeless of machine learning such as Random Forest, Extremely Random Trees, and Gradient Boosting Trees
3- The GBRT has shown the less error rate and can be used as the best model

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author statement:

Yaoyang Liu: conception, experimental design, carrying out measurements and manuscript composition, writing

Drai Ahmed Smait: writing, investigation, experimental design, carrying out measurements, modelling and analysis

Abbas Yaseen Naser: resources, experimental design, carrying out measurements, writing, analysis

Farag M. A. Altalbawy: investigtaion, experimental design, carrying out measurements and writing, formal analysis

Hala Bahri: writing and editing, investigation, validation

Ali Abdul Kadhim Ruhaima: writing and editing, investigation, validation, resources

Thura Zayad Fathallah: writing and editing, investigation, validation, resources, analysis

Salema K. Hadrawi: writing and editing, investigation, resources, analysis

Refad E. Alsaddon: writing and editing, investigation, resources, analysis

Abdullah Alshetaili: writing and editing, investigation, resources, analysis

Amal M. Alsubaiyel: writing and editing, investigation, validation, resources, analysis, supervision