



OPEN Intelligence computational analysis of letrozole solubility in supercritical solvent via machine learning models

Mohammed Alqarni¹✉, Amal Adnan Ashour², Alaa Shafie³, Ali Alqarni², Mohammed Fareed Felemban², Bandar Saud Shukr⁴, Mohammed Abdullah Alzubaidi⁴ & Fahad Saeed Algahtani⁵

Supercritical fluids (SCFs) can be used to prepare drugs nanoparticles with improved solubility. SCFs have shown superior advantages in pharmaceutical industry as an environmentally friendly alternative to toxic/harmful organic solvents. They possess gas-like transport characteristics and liquid-like solvation power for solutes. Evaluation of chemotherapeutic drugs' solubility in supercritical carbon dioxide (SCCO₂) has been recently an attractive subject for developing this method in pharmaceutical sector. To reach this purpose, the utilization of accurate models is of great necessity to estimate experimental-based solubility data. In this paper, the authors tried to employ machine learning (ML) approaches to estimate the solubility of Letrozole (LET) drug as chemotherapeutic agent and correlate its values in wide ranges of temperature and pressure. To do this, PAR (Passive Aggressive Regression), RF (Random Forest), and RBF-SVM are the models used (Support Vector Machine with RBF kernel). These models optimized in terms of their hyper-parameters using GA algorithm. The optimized PAR, RF, RBF-SVM models obtained coefficients of determination (R-squared) of 0.8277, 0.9534, and 0.9947. Also, the MSE error rate of the models are 0.1342, 0.0305, and 0.0045, in the same order. The final result of the evaluations shows the optimized RBF-SVM model as the most appropriate model in this research. The model exhibits a maximum prediction error of 0.1289.

Keywords Letrozole, Solubility, Supercritical processing, Machine learning

In the recent years, major efforts of different international research and development (R&D) groups in the pharmaceutical industry have been focused on the development of novel, efficacious and affordable therapeutic drugs with optimum safety, minimum side effects and acceptable biological/physicochemical characteristics^{1,2}. Because of the availability of great numbers of active pharmaceutical ingredients (APIs) in solid state, one of the traditional drugs processing techniques is based on their recrystallization applying organic solvents as reaction media³. Despite the prevalence of application, the emergence of some drawbacks like environmental detriments, toxicity during application and wide particle size distribution has significantly restricted the use of organic solvents in drug industry^{4,5}. To mitigate the negative effects of organic solvents, Supercritical fluids (SCFs) have been welcomed at different scientific areas such as solid drugs' extraction and nanonization of solid dosage forms⁶⁻⁸. Industrial-based use of SCFs (especially SCCO₂) is due to their precious advantages such as great potential of application in harsh operational conditions, eco-friendliness and proper efficiency^{9,10}. SCCO₂ occurs at the pressures/temperatures higher than the critical point of CO₂, which causes its properties to be between the liquid state and gas state¹¹⁻¹³. SCFs have indicated great capabilities in various fields including extraction, food science, nanotechnology, and drug preparation¹⁴⁻¹⁷.

Recently, optimization of mathematical models to determine the solubility of various kind of drugs is one of the important tasks of scientists to estimate the behavior of drug release at different temperatures and

¹Department of Pharmaceutical Chemistry, College of Pharmacy, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia. ²Department of Oral & Maxillofacial Surgery and Diagnostic Sciences, Faculty of Dentistry, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia. ³Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Taif University, P.O.Box 11099, Taif 21944, Saudi Arabia. ⁴Department of Preventive Dentistry, Faculty of Dentistry, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia. ⁵Department of Restorative Dental Science, Faculty of Dentistry, Taif University, Taif 21944, Saudi Arabia. ✉email: m.aalqarni@tu.edu.sa

pressures^{18–20}. Letrozole (Femara[®]) with the chemical formulation of C₁₇H₁₁N₅ is a well-known anti-neoplastic orally-administered drug, which is extensively applied to the second-line treatment of women suffering from certain types of breast cancer^{21,22}.

Machine learning (ML) has garnered significant momentum across various scientific disciplines, gradually supplanting traditional computing methods in an ever-expanding array of scientific domains by assuming their functionalities²³. Here, we used machine learning methods to estimate solubility of LET drug. The employed models are RF (Random Forest), PAR (Passive Aggressive Regression), and RBF-SVM (Support Vector Machine with RBF kernel). Passive Aggressive Regression (PAR) is among the models used in this study that belongs to online learning approach. Unlike batch learning techniques, which develops the estimator by training on the full training data simultaneously, online machine learning allows the best predictor for future data to be updated at each step as new data becomes available²⁴. In the ensemble learning model Random Forest (RF), voting is employed to improve the efficacy of learners featuring diverse base trees²⁵. A random forest is popular because it predicts numerous events with minimal parameters. This approach can handle both a high-dimensional feature space and tiny examples accurately. They could handle big systems in the real world because they are parallelizable²⁶. Support SVMs are well-known supervised models that can be used for both classification and regression tasks. Because processing takes time, it is used for smaller data sets. This method is based on the idea of finding a hyperplane that best separates inputs in distinct locations^{27–29}. Here we used SVM model with the RBF kernel (RBF-SVM). These models optimized in terms of their hyper-parameters using GA algorithm as an innovation. The purpose of the current research article is to optimize three novel approaches (PAR, RF RBF-SVM) using artificial intelligence technique to estimate and develop the solubility value of Letrozole anti-cancer drug through the SCCO₂. The models and their combination with optimizer (GA algorithm) are developed and implemented for the first time to correlate solubility data of Letrozole.

Data set for machine learning computing

For this work, there are 45 rows of data about LET (Letrozole) drug solubility that can be split down as follows: The inputs consist of numerical values, and the resulting output is likewise a numerical value. The data used in this investigation, which was acquired from references, is shown in Table 1³⁰. All data analytics and fitting in this study are performed using Python software.

Methodology of intelligence computing PAR (passive aggressive regression) method

To train a machine learning model online, we feed it instances one at a time or in small batches known as mini batch. As a result, problems with continuously streaming data are well-suited to the passive-aggressive approach³¹. Although this model is often employed for use in large-scale tasks, it can be used to much smaller data sets and still shows remarkable robustness. Passive aggressive regressors are similar to Perceptron when a learning rate is not needed. They differ from the Perceptron in that they have a regularization parameter C ^{32,33}. Classification with the squared hinge or raw hinge loss functions, and regression with the insensitive or squared insensitive loss functions, are all possible with this model as defined here³¹:

$$L = \begin{cases} 0 & \text{if } |y_i - y_i| - \epsilon \leq 0 \\ |y_i - y_i| - \epsilon & \text{otherwise} \end{cases}$$

Random forest

The Random Forest (RF) method combines the results of numerous Decision Tree (DT) models to approximate the desired value for a set of data points^{34,35}. Upon receiving an input data point containing attribute values, like x , the Random Forest algorithm develops K individual trees and then computes the average of their outputs to generate a final prediction. Following is the formula for the RF model with K trees $T(x)$ ²⁶.

P (MPa)	S ($\times 10$ g L ⁻¹)				
	T = 308 (K)	T = 318 (K)	T = 328 (K)	T = 338 (K)	T = 348 (K)
12.2	0.11	0.09	0.04	0.03	0.03
15.2	0.19	0.18	0.17	0.17	0.15
18.2	0.19	0.33	0.69	0.54	0.1
21.3	0.24	0.48	0.63	0.92	1.03
24.3	0.25	0.51	0.81	1.27	1.55
27.4	0.28	0.56	1.09	1.68	2.21
30.4	0.33	0.63	1.12	2.21	3.39
33.4	0.33	0.66	1.33	2.57	4.04
35.5	0.35	0.73	1.45	2.79	4.36

Table 1. Used data for intelligence computing³⁰.

$$\hat{f}_{rf}^K(x) = \frac{1}{K} \sum_{k=1}^K T(x)$$

Random Forest utilizes bagging method to enhance the diversity of the decision tree model, enabling them to be developed based on varied training set. Using the method, the entire training dataset undergoes random resampling, and the replacement data is retained. Consequently, specific data may undergo frequent utilization, while others might be employed only once. This particular strategy enhances the precision of predictions and enhances overall stability by demonstrating increased resilience to minor fluctuations in input data. In contrast, trees generated by RF use a randomly selected subset of features to determine the optimal design. This approach diminishes the overall strength of the forest, concurrently reducing the correlation among individual trees and thereby minimizing generalization error. Furthermore, as they develop, RF estimator trees do not need to be clipped, making them more computationally efficient^{34,36,37}. Also, the out-of-bag subset is generated as an independent subset, commonly referred to as OOB. It comprises samples from the bagging process that were not picked for training the k -th tree. The k th tree can measure performance based on these OOB elements^{38,39}.

Support vector machine (SVM) with RBF Kernel

In the SVM method, independent variable x aids in the estimation of the dependent variable y , according to⁴⁰. Like in other regression scenarios, the given function determined the correlation between x and y ⁴¹.

$$f(x) = f(x) \cdot wT \cdot \phi(x) + b$$

$$f(y) = f(x) + \text{noise}$$

In recent equations, ϕ stands for a kernel that accepts input data and transforms it into the required shape. SVM methods use various types of kernel functions. Polynomial, linear, sigmoid, and RBF are some examples (Here the RBF is used). w is the vector coefficient, b is a constant, and b and w are the regression function constraints. Error tolerance, on the other hand, creates noise (e). During the Support Vector Machine model's training, consecutive minimization association process of the error function can be obtained. Based on the error function, there are two types of SVM models: e-SVM and t-SVM⁴².

The RBF function is defined as⁴³:

$$K(x, x_i) = \exp\left(\frac{-|x - x_i|^2}{\sigma^2}\right)$$

This function was used as a kernel in this study.

Results and discussions

As highlighted in the introduction, the introduced models have been developed through the help of genetic algorithm and the final statistical results of these models are displayed in Table 2. Based on this table, the best model is clearly the RBF-SVM model, and the RF model is on the second rank, and the PAR model is in the third place with acceptable air performance.

Figures 1, 2 and 3 also confirm this fact. In these figures, the experimental values have been checked with the values predicted by the models. In these three figures, there is a $Y = X$ line, which represents the experimental data. The blue points illustrate the estimated values in the training and the red crosses are the predicted solubilities of test subset.

Although the RBF-SVM model was chosen as the best model and other analyzes were performed based on this model, but due to the acceptable performance of all three models, Figs. 4, 5 and 6 showcase the decision level diagrams in a three-dimensional format for each of the three models. As presented, the solubility value of Letrozole improves as the pressure does. It can be attributed to this fact that the increment of pressure increases the CO_2 and thus, the solvent power improves. Increment of solvation strength takes place due to the decrement of intermolecular distance and as the result, enhancement of solute-solvent interactions. The impact of temperature on the solubility behavior is more complex. Based on the results, at the pressures greater than the cross over pressure, the positive impact of vapor pressure increment is higher than the destructive effect of solvent density decrement. Moreover, an enhancement of T at these points improves the solubility of Letrozole. For the points less than the cross over pressure, the negative effect of decreasing in density overcomes the positive influence of vapor pressure increment⁴⁴. Therefore, an enhancement in temperature declines the Letrozole solubility.

As mentioned before, the RBF-SVM model has the best performance among these three models. Figure 7, which shows the residual of this model, also confirms the good efficiency of this approach. In addition, the RF

Models	Train R^2	Test R^2	MAE	MSE	MAPE	Max error
PAR	0.9136	0.8277	0.2697	0.1342	0.2990	0.7611
RF	0.9891	0.9534	0.1132	0.0305	0.1855	0.4659
RBF SVM	0.9944	0.9947	0.0561	0.0045	0.0888	0.1289

Table 2. Final model results.

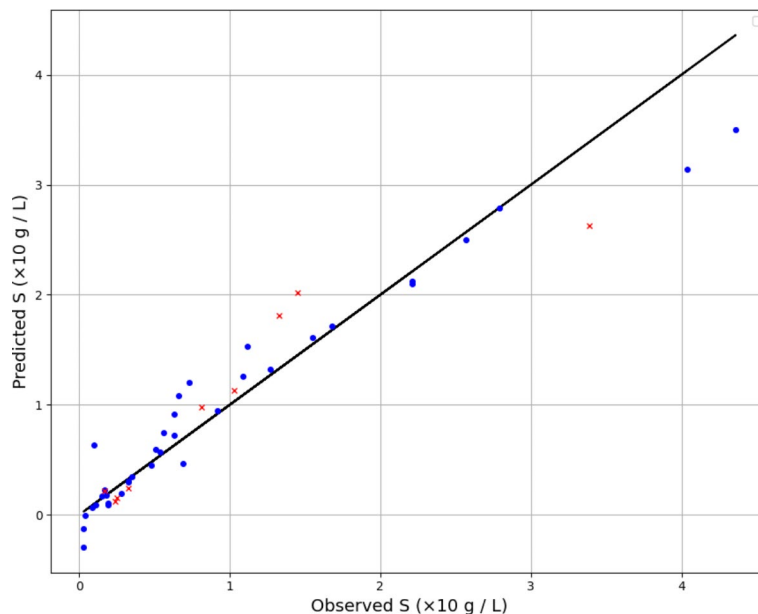


Fig. 1. Predicted vs. observed data (PAR approach).

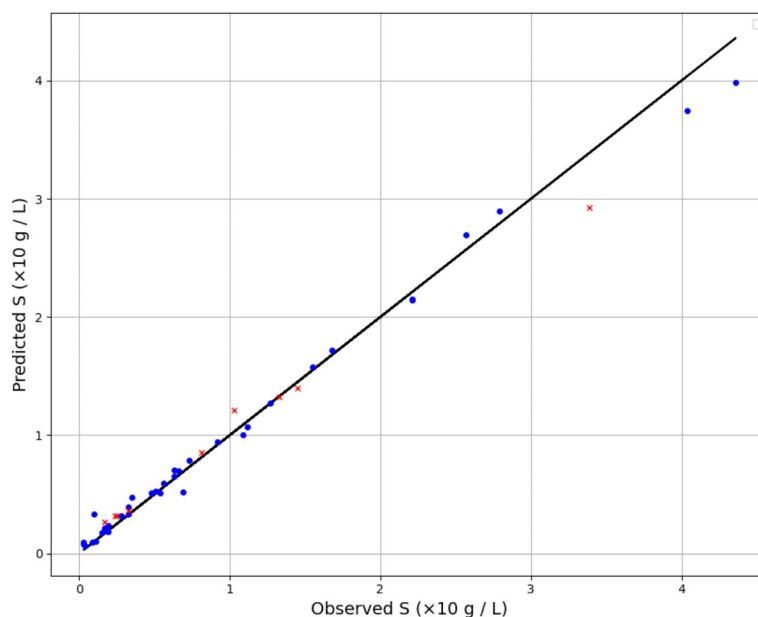


Fig. 2. Predicted vs. observed data (RF approach).

approach has the second performance by a short distance, so Fig. 8 also shows the importance of the features with the help of the Random Forest model. Finally, the influence of individual features on the solubility are illustrated in Figs. 9 and 10.

Conclusion

Development of promising techniques to improve the solubility, dissolution ratio and bioavailability of oral-dosage chemotherapeutic medicines is still a challenge in the drug manufacturing industry. To reach these purposes, some predictive models should be developed to precisely optimize the drug solubility. In this research article, ML-based predictive models were employed to develop the solubility value of Letrozole anti-cancer drug through the SCCO_2 at different ranges of pressure and temperatures. The employed models are PAR (Passive Aggressive Regression), RF (Random Forest), and RBF-SVM (Support Vector Machine with RBF kernel). The GA algorithm was used to optimize these models' hyper-parameters. The coefficients of determination (R^2) for the optimized PAR, RF, and RBF-SVM approaches are 0.8277, 0.9534, and 0.9947. In addition, the models'

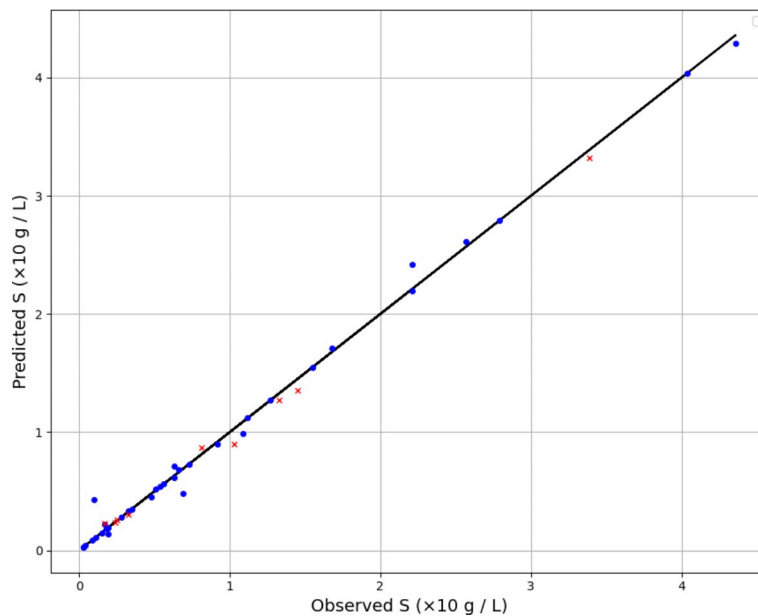


Fig. 3. Predicted vs. observed data (RBF-SVM approach).

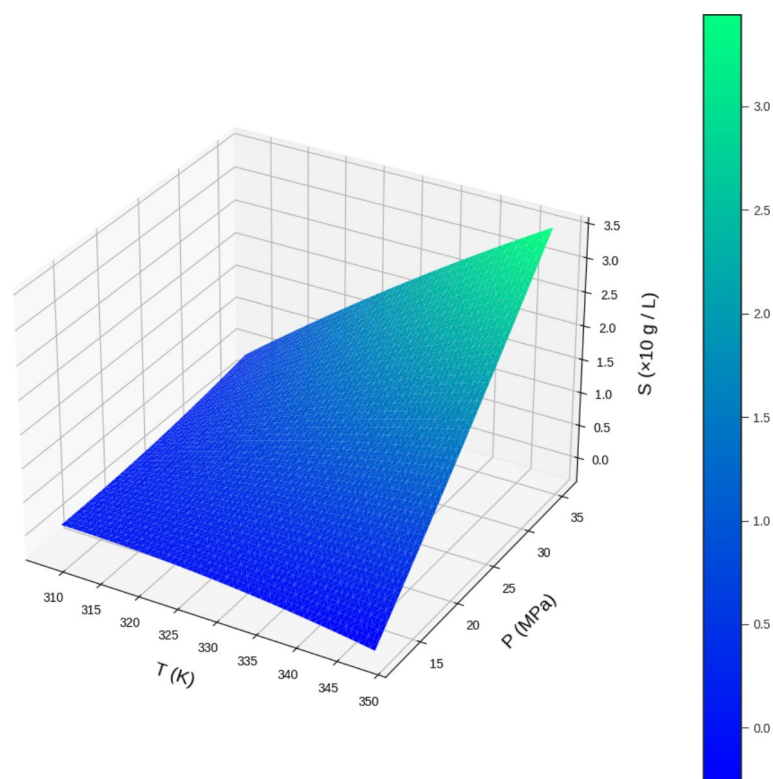


Fig. 4. Prediction 3D surface (PAR MODEL).

MSE error rates are 0.1342, 0.0305, and 0.0045, in that order. The validated outputs exhibited that the optimized RBF-SVM approach is the best fit for this research. Using this model, the maximum prediction error is 0.1289.

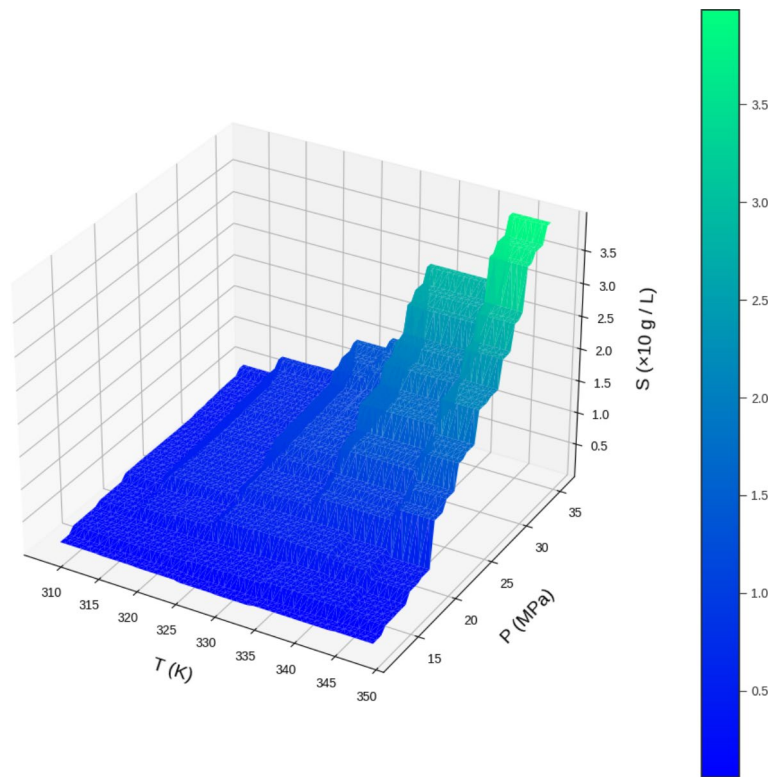


Fig. 5. Prediction 3D surface (RF MODEL).

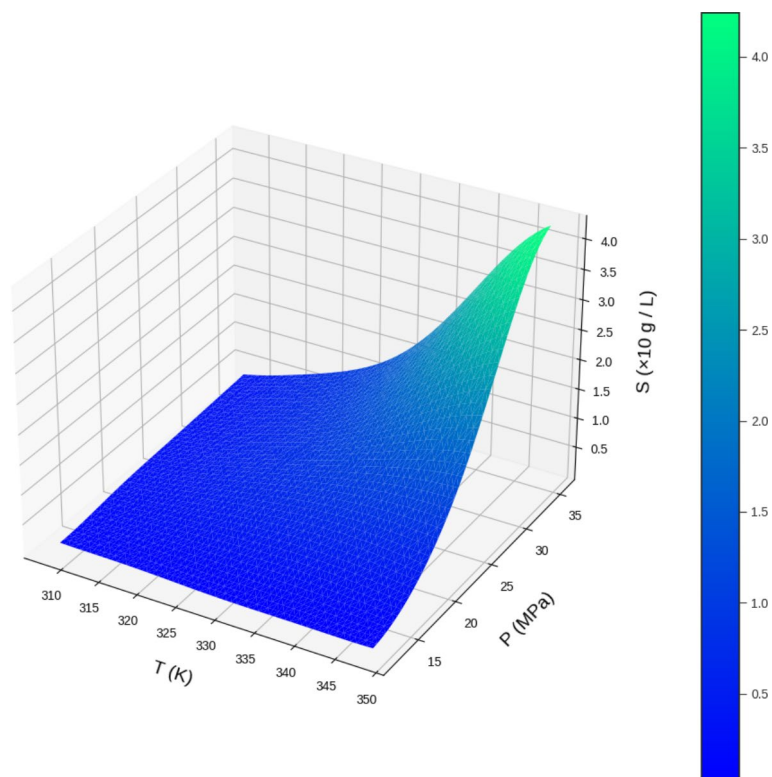


Fig. 6. Prediction 3D surface (RBF-SVM MODEL).

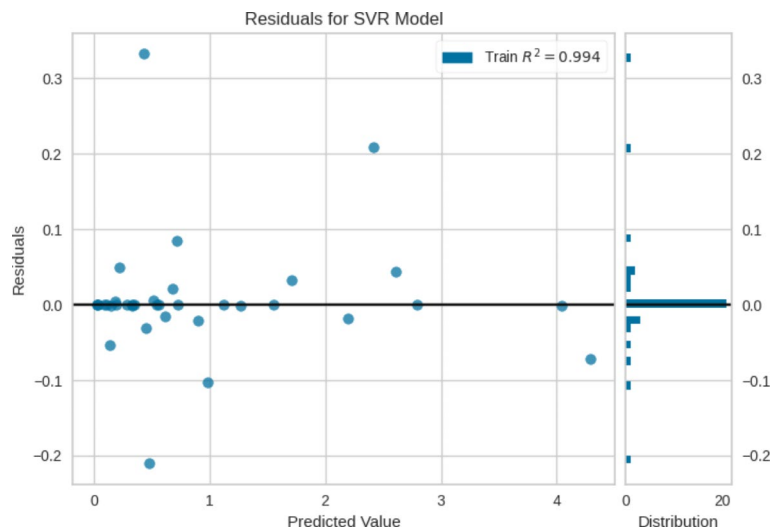


Fig. 7. Final residuals of RBF-SVM model.

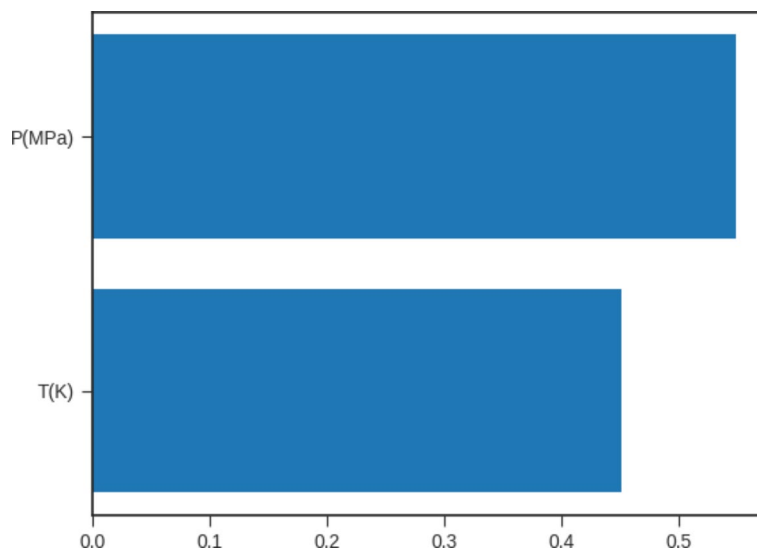


Fig. 8. Feature importance using RF.

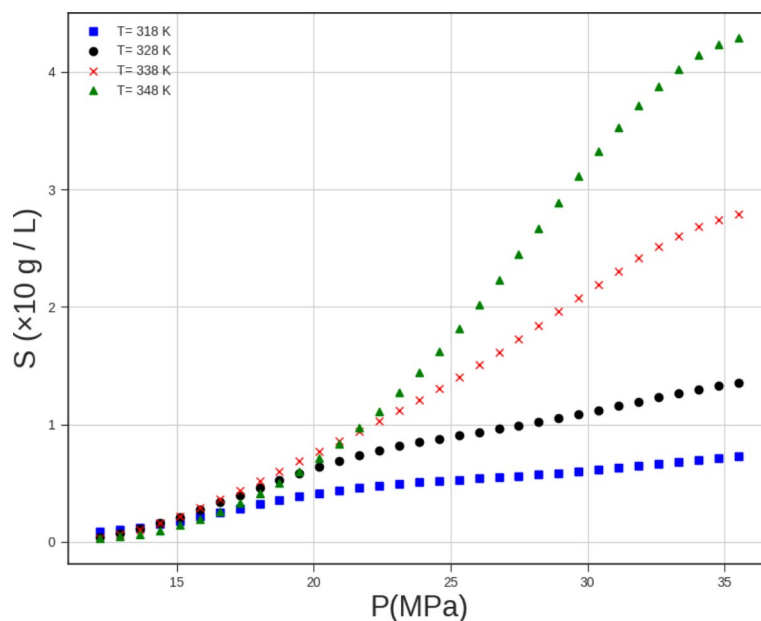


Fig. 9. Impact of pressure on solubility at various temperature.

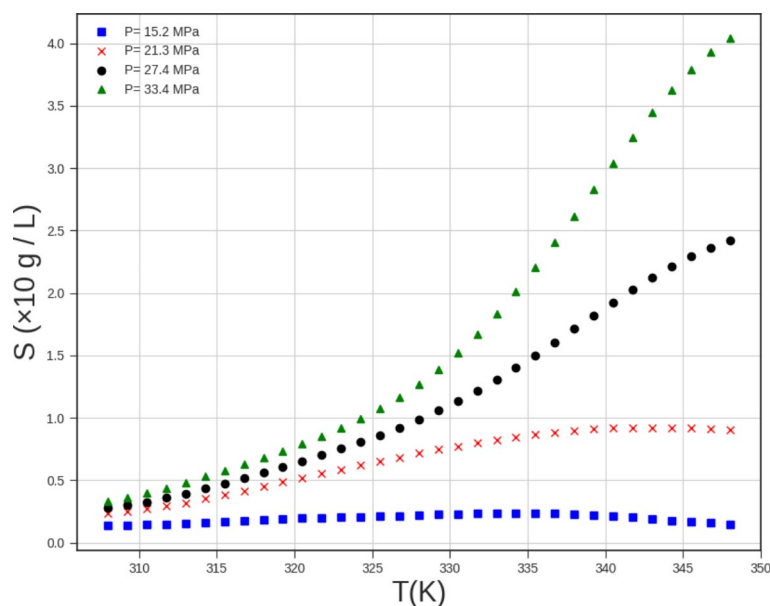


Fig. 10. Impact of temperature on solubility at various Pressure.

Data availability

The data supporting this study are available when reasonably requested from the corresponding author.

Received: 2 August 2024; Accepted: 12 September 2024

Published online: 17 September 2024

References

1. Bagheri, H. et al. Supercritical carbon dioxide utilization in drug delivery: experimental study and modeling of Paracetamol solubility. *Eur. J. Pharm. Sci.* **177**, 106273 (2022).
2. Padrela, L. et al. Supercritical carbon dioxide-based technologies for the production of drug nanoparticles/nanocrystals—a comprehensive review. *Adv. Drug Deliv. Rev.* **131**, 22–78 (2018).
3. Molani, S., Madadi, M. & Wilkes, W. A partially observable Markov chain framework to estimate overdiagnosis risk in breast cancer screening: incorporating uncertainty in patients adherence behaviors. *Omega* **89**, 40–53 (2019).

4. Yan, J. et al. Chiral protein supraparticles for tumor suppression and synergistic immunotherapy: an enabling strategy for bioactive supramolecular chirality construction. *Nano Lett.* **20** (8), 5844–5852 (2020).
5. Dikmen, G., Genç, L. & Güney, G. Advantage and disadvantage in drug delivery systems. *J. Mater. Sci. Eng.* **5** (4), 468 (2011).
6. Sodeifian, G. & Sajadian, S. A. Solubility measurement and preparation of nanoparticles of an anticancer drug (letrozole) using rapid expansion of supercritical solutions with solid cosolvent (RESS-SC). *J. Supercrit. Fluids* **133**, 239–252 (2018).
7. Rojas, A. et al. Improving and measuring the solubility of favipiravir and montelukast in SC-CO₂ with ethanol projecting their nanonization. *RSC Adv.* **13** (48), 34210–34223 (2023).
8. Askarizadeh, M. et al. Binary and ternary approach of solubility of Rivaroxaban for preparation of developed nano drug using supercritical fluid. *Arab. J. Chem.* **17** (4), 105707 (2024).
9. Khandare, K. & Goswami, S. Extraction of kaemferol from *Moringa oleifera* using CO₂ supercritical fluid extraction: a green technology. *AIJR Abstr.* **66** (2021).
10. Mihalcea, L. et al. CO₂ supercritical fluid extraction of oleoresins from Sea Buckthorn Pomace: evidence of advanced bioactive profile and selected functionality. *Antioxidants* **10** (11), 1681 (2021).
11. Long, B., Ryan, K. M. & Padrela, L. From batch to continuous—new opportunities for supercritical CO₂ technology in pharmaceutical manufacturing. *Eur. J. Pharm. Sci.* **137**, 104971 (2019).
12. Abdelbasset, W. K. et al. Modeling and computational study on prediction of pharmaceutical solubility in supercritical CO₂ for manufacture of nanomedicine for enhanced bioavailability. *J. Mol. Liq.* **359**, 119306 (2022).
13. Zhuang, W. et al. Ionic liquids in pharmaceutical industry: a systematic review on applications and future perspectives. *J. Mol. Liq.* **349**, 118145 (2022).
14. Sodeifian, G. & Sajadian, S. A. Investigation of essential oil extraction and antioxidant activity of *Echinophora platyloba* DC. Using supercritical carbon dioxide. *J. Supercrit. Fluids* **121**, 52–62 (2017).
15. Sodeifian, G., Azizi, J. & Ghoreishi, S. M. Response surface optimization of *Smyrniun cordifolium* Boiss (SCB) oil extraction via supercritical carbon dioxide. *J. Supercrit. Fluids* **95**, 1–7 (2014).
16. Sodeifian, G., Sajadian, S. A. & Saadati Ardestani, N. Supercritical fluid extraction of omega-3 from *Dracocephalum Kotschy* seed oil: process optimization and oil properties. *J. Supercrit. Fluids* **119**, 139–149 (2017).
17. Ameri, A., Sodeifian, G. & Sajadian, S. A. Lansoprazole loading of polymers by supercritical carbon dioxide impregnation: impacts of process parameters. *J. Supercrit. Fluids* **164**, 104892 (2020).
18. Wang, W. et al. *Interdisciplinary Evolution of the Machine Brain* 119–145 (Springer, 2021).
19. Jin, K. et al. Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. *Acta Ophthalmol.* **100** (2), e512–e520 (2022).
20. Sodeifian, G. et al. A comprehensive comparison among four different approaches for predicting the solubility of pharmaceutical solid compounds in supercritical carbon dioxide. *Korean J. Chem. Eng.* **35** (10), 2097–2116 (2018).
21. Bryant, J. & Wolmark, N. *Letrozole After Tamoxifen for Breast Cancer—What is the Price of Success?* 1855–1857 (Mass Medical Soc, 2003).
22. Simpson, D., Curran, M. P. & Perry, C. M. Letrozole drugs **64** (11), 1213–1230 (2004).
23. Alpaydin, E. *Introduction to Machine Learning* (MIT Press, 2020).
24. Fontenla-Romero, Ó. et al. Online machine learning. In *Efficiency and Scalability Methods for Computational Intellect* 27–54 (IGI Global, 2013).
25. Jiang, R. et al. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinform.* **10** (1), 1–12 (2009).
26. Seyghaly, R. et al. Interference recognition for fog enabled IoT architecture using a novel tree-based method. In *IEEE International Conference on Omni-Layer Intelligent Systems (COINS)* (IEEE Computer Society, 2022).
27. Meyer, D., Leisch, F. & Hornik, K. The support vector machine under test. *Neurocomputing* **55** (1–2), 169–186 (2003).
28. Mangasarian, O. L. & Musicant, D. R. Robust linear and support vector regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (9), 950–955 (2000).
29. Alamri, A. & Alafnan, A. Artificial intelligence optimization of Alendronate solubility in CO₂ supercritical system: Computational modeling and predictive simulation. *Ain Shams Eng. J.* **15** (9), 102905 (2024).
30. Hojjati, M. et al. Supercritical CO₂ and highly selective aromatase inhibitors: Experimental solubility and empirical data correlation. *J. Supercrit. Fluids* **50** (3), 203–209 (2009).
31. Crammer, K. et al. *Online Passive Aggressive Algorithms* (2006).
32. Yin, G. et al. Machine learning method for simulation of adsorption separation: comparisons of model's performance in predicting equilibrium concentrations. *Arab. J. Chem.* **15** (3), 103612 (2022).
33. Adun, H. et al. Impact of data processing and robust machine learning process on accurate estimation of specific heat capacity property in energy storage applications. *J. Energy Storage* **55**, 105359 (2022).
34. Breiman, L. Random forests. *Mach. Learn.* **45** (1), 5–32 (2001).
35. Rodriguez-Galiano, V. F. et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **67**, 93–104 (2012).
36. Almunirawi, K. M. & Maghari, A. Y. A comparative study on serial decision tree classification algorithms in text mining. *Int. J. Intell. Comput. Res.* **7** (4) (2016).
37. Verikas, A., Gelzinis, A. & Bacauskiene, M. Mining data with random forests: a survey and results of new tests. *Pattern Recogn.* **44** (2), 330–349 (2011).
38. Peters, J. et al. Random forests as a tool for ecohydrological distribution modelling. *Ecol. Model.* **207** (2–4), 304–318 (2007).
39. Liu, Z. et al. Development of compositional-based models for prediction of heavy crude oil viscosity: application in reservoir simulations. *J. Mol. Liq.* **389**, 122918 (2023).
40. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, 1999).
41. Waqas, M. et al. Evaluating the performance of different artificial intelligence techniques for forecasting: rainfall and runoff prospective. *Weather Forecast.* **23** (2021).
42. Noori, R. et al. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.* **401** (3–4), 177–189 (2011).
43. Kuo, B. C. et al. A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **7** (1), 317–326 (2013).
44. Liu, Y. et al. Optimization and validation of drug solubility by development of advanced artificial intelligence models. *J. Mol. Liq.* **372**, 121113 (2023).

Acknowledgements

The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-122).

Author contributions

Mohammed Alqarni: Validation, Formal analysis, Investigation, Writing—Original Draft, Visualization, Amal

Adnan Ashour: Conceptualization, Investigation, Writing—Original Draft, Visualization, Alaa Shafie: Conceptualization, Formal analysis, Methodology, Writing—Original Draft, Visualization, Ali Alqarni: Conceptualization, Formal analysis, Investigation, Writing—Review & Editing, Visualization, Mohammed Fareed Felemban, Bandar Saud Shukr: Conceptualization, Formal analysis, Investigation, Writing—Original Draft, Visualization, Mohammed Abdullah Alzubaidi: Formal analysis, Investigation, Writing—Review & Editing, Visualization, Fahad Saeed Algahtani: Conceptualization, Formal analysis, Investigation, Writing - Original Draft, Validation.

Funding

This research was funded by Taif University, Saudi Arabia, Project No. (TU-DSPP-2024-122).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024