scientific reports

Check for updates

OPEN

Optimization of drug solubility inside the supercritical CO_2 system via numerical simulation based on artificial intelligence approach

Meixiuli Li[1], Wenyan Jiang[2]✉, Shuang Zhao[1], Kai Huang[1] & Dongxiu Liu[3]

In this research paper, we explored the predictive capabilities of three different models of Polynomial Regression (PR), Extreme Gradient Boosting (XGB), and LASSO to estimate the density of supercritical carbon dioxide (SC-CO_2) and the solubility of niflumic acid as functions of the input variables of temperature and pressure. The optimization of hyperparameters for these models is achieved using the innovative Barnacles Mating Optimizer (BMO) algorithm. For SC-CO_2 density estimation, PR exhibits remarkable accuracy, showing an R-squared value of 0.99207 for data fitting. XGB performs admirably with an R^2 of 0.92673, while LASSO model demonstrates good predictive ability, showing an R^2 of 0.81917. Furthermore, we assess the models' performance in predicting the solubility of niflumic acid. PR exhibits excellent predictive capabilities with an R^2 of 0.96949. XGB also delivers strong performance, yielding an R-squared score of 0.92961. LASSO performs well, achieving an R-squared score of 0.82094. The results indicated promising performance of machine learning models and optimizer in estimating drug solubility in supercritical CO_2 as the solvent applicable for pharmaceutical industry.

Keywords Polynomial regression, Extreme Gradient Boosting, LASSO, Drug solubility, Supercritical CO_2

Due to the inherent hydrophobic nature of pharmaceuticals, most drugs are poor soluble or practically insoluble in aqueous solutions which makes the commercialization stage of some drugs impossible. Another problem with low solubility of drugs is that they need to be taken at higher dosage to achieve the therapeutic effects. So, their efficacy is low, and it should be enhanced. Different techniques can help improve the solubility of drug substances which rely on increasing the solubility through chemical and physical methods[1–3]. For instance, ball milling is a facile method to reduce the size of drug particles and increase their solubility due to the smaller size[4]. The method of pharmaceutical cocrystallization is another approach to enhance drugs solubility which is based on molecular interactions between drug and a coformer to build a combination of species with enhanced solubility in aqueous media[5].

For commercialization and application of a wide range of medications, more attractive processes and techniques are needed. For instance, the method of drugs nanonization via supercritical fluids has been assessed and studied recently to enhance the solubility of medications by production of nanosized drugs particles. The method is attractive owing to the utilization of supercritical fluids such as CO_2 (SC-CO_2) which are green solvents[6]. There are different steps involved in this process among which the drug solubility in the supercritical fluid is the most important one as the process efficiency is determined by the solubility of drug. Furthermore, due to the variation of pressure and temperature, the solubility must be evaluated as a function of these parameters. Some techniques have been proposed to evaluate drugs solubility in SC-CO_2such as thermodynamics and machine learning. Despite the physical basis of thermodynamic models in evaluating drug solubility[7,8], machine learning models have offered higher precision in estimation of drugs solubility in supercritical fluids[9–12].

Machine learning (ML) models have emerged as effective methods for data analysis and predictive modeling across a wide range of domains. Over the previous decade, there has been significant progress in the field of ML, as evidenced by the development of a plethora of algorithms and models aimed at addressing a wide range of applications including drug development[13,14]. Several ML models have been already reported for correlation of drugs solubility, but new models should be customized for a new drug to make the generalized framework

[1]Department of Human Anatomy and Embryology, Pu Ai Medical School, Shaoyang University, Shaoyang 422000, Hunan, China. [2]The Second Affiliated Hospital of Shaoyang University, Shaoyang University, Shaoyang 422000, Hunan, China. [3]Shashi Town Health Center, Shaodong 422813, Hunan, China. ✉email: 17670918058@163.com

for analysis of drugs solubility in SC-CO_2 solvent. Here, for the first time Polynomial Regression (PR), Extreme Gradient Boosting (XGB), and LASSO are developed and optimized for estimation of niflumic acid solubility in SC-CO_2. The Barnacles Mating Optimizer (BMO) was used to train and optimize these models. Indeed, utilization of ML models, optimization, and implementation for niflumic acid is carried out for the first time in this research.

In the field of linear regression analysis, LASSO regression is a powerful and widely used technique. It effectively addresses multicollinearity, overfitting, and high-dimensional data, balancing model interpretability and predictive accuracy. LASSO promotes feature selection and regularization, which helps to develop more robust and parsimonious models in a variety of domains[15,16]. PR is a useful extension of linear regression that allows us to model nonlinear relationships between variables. This technique can capture complex patterns in data and provide valuable insights into the underlying relationships by introducing higher-order polynomial terms[17]. The XGBoost regression algorithm is a commonly used machine learning algorithm that is well-known for its high predictive accuracy and efficient handling of complex datasets. XGBoost produces robust and interpretable regression models for a variety of applications by combining gradient boosting, regularization, and decision trees. Its ability to handle missing data, identify feature importance, and resist overfitting adds to its appeal among data scientists and practitioners seeking accurate and reliable predictive modeling[18,19].

Data of solubility

The dataset utilized in this work is collected from a previous work[20] and contains four distinct variables for drug solubility in supercritical CO_2, namely Temperature, Pressure, Solvent Density, and Solubility of niflumic acid. The experimental conditions are represented by the Temperature and Pressure values, while the corresponding measurements are provided by the Solvent Density and Solubility of niflumic acid as reported in[20]. As such, two inputs and two outputs are assumed for building the machine learning models in this study.

Methodology

Barnacles mating optimizer (BMO)

BMO draws inspiration from the mating behavior of barnacles which is used for tuning models in this work. These microorganisms are regarded as promising candidates (combinations of hyperparameters in this study) in this algorithm[21]. The BMO process comprises two primary stages, namely selection and reproduction. Two parent barnacles are chosen for the selection phase according to the length of their penises (*pl*).

The Hardy-Weinberg principle is used by the algorithm to generate offspring during the reproduction stage. If the *pl* of the father's barnacle falls within the selection range of the parent barnacles, the father inherits *p%* of the characteristics and the mother inherits *(1-p)%*. If the father's *pl* is outside the range of acceptable mutations, a new generation is generated by modifying only the maternal traits. This approach promotes exploitation when the father's *pl*is within the range and exploration when it is not[22].

The formulation for generating offspring from the parents' mating process is expressed through the following Eqs[21,22]:

$$x_i^{N_new} = px_{barnacle_d}^N + qx_{barnacle_m}^N$$

In this process, the generation of offspring relies on two random numbers, *p* and *q*, both falling within the range of *[0, 1]*. Here, barnacled represents the solution for the father, and $barnacle_m$ represents the solution for the mother. If barnacled chooses barnacle8, it exceeds the cap, leading to the termination of the usual mating process.

Instead, the algorithm employs a method called "sperm cast," a term coined in BMO, to generate the offspring. This approach facilitates exploration during the mating process[22]:

$$x_i^{N_new} = rand() \times x_{barnacle_m}^n$$

The function *rand()* generates a random number within the interval [0, 1].

LASSO regression

LASSO (Least Absolute Shrinkage and Selection Operator) is an advanced statistical technique used in linear regression applications. It was introduced as a method to handle multicollinearity and perform feature selection by imposing a penalty on the absolute values of the regression coefficients. This model has gained popularity due to its ability to effectively handle high-dimensional datasets and produce interpretable and sparse models[16,23].

The primary objective of LASSO regression is to determine the best linear model by minimizing the sum of squared residuals while simultaneously shrinking the less informative coefficients to zero. This encourages the selection of the most relevant features and avoids overfitting, leading to a more robust and generalizable model.

Let's consider a linear regression problem with *n* observations and *p* predictors. The model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

where:

- - *y* stands for the dependent variable,
- - β_0 indicates the intercept,
- - x_i's are the predictors,

- - β_i's are the coefficients, and
- - ϵ represents the error.

The LASSO regression optimizes the following objective function[24]:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The symbol λ represents the regularization factor. As λ increases, the penalty for non-zero coefficients strengthen, leading to more shrinkage and feature selection.

Extreme gradient boosting (xgboost)

XGBoost has gained widespread acclaim for its high predictive performance and versatility across a wide range of domains. As an ensemble learning technique, XGBoost combines the strengths of gradient boosting and regularization to deliver robust and accurate regression models[19,25].

The primary objective of XGBoost regression is to create an optimized regression model that can effectively predict continuous numeric values. By employing a combination of weak learners (Decision Trees in this study), typically decision trees, XGBoost progressively improves its predictive capability through iterative boosting. It aims to minimize the overall prediction error and deliver superior results compared to traditional gradient boosting algorithms. A Flowchart for overall process of XGBoost is displayed in Fig. 1[26].

Polynomial regression (PR)

PR method allows for the modeling of nonlinear relationships between the inputs and outputs for complicated tasks. By introducing polynomial terms, this technique can capture complex patterns where linear models fail. In this model description, we explore the key concepts and benefits of polynomial regression[27,28]. In this regression method, PR model is employed to fit a polynomial function to the data in order to approximate the underlying relationship between the variables. Polynomial regression can capture curved and nonlinear trends in the data[29].

The PR of order *d*is given by[12]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \beta_{p+1} x_1^2 + \beta_{p+2} x_1 x_2 + \ldots + \beta_{p+n} x_p^d + \epsilon$$

where *d* is the PR order, x_i^d represents the *d*-th power of the *i*-th predictor, and β_{p+1} to β_{p+n} are the additional parameters to be estimated.

For the ML modeling and optimization tasks, *Python* software was used along with machine learning, optimization, and plotting libraries.

Results and discussion

The results of three models in estimating the solubility of niflumic acid and corresponding density of solvent using temperature and pressure as inputs are presented in this section. Indeed, both responses have been modeled and their values are compared by three models to find out the accuracy of optimized models in this study. The results of analyses for all models and responses are listed in Tables 1 and 2. Three important criteria have been considered for comparison including R^2 (Coefficient of Determination), RMSE (Root mean square error), and Maximum Error.

From the results listed in Tables 1 and 2, it is evident that Polynomial Regression (PR) consistently outperforms the other ML models in predicting both outputs, i.e., SC-CO_2 density and niflumic acid solubility. The comparison of real and predicted values for both outputs is shown in Figs. 2 and 3 which illustrates the dataset for training and testing. PR achieves remarkable accuracy, with high R^2 scores of 0.992 and 0.969 for density and solubility, respectively, and RMSE values of 12.203 and 0.256. XGB also demonstrates good predictive performance, with R-squared scores of 0.927 and 0.930, and RMSE values of 28.623 and 0.286. LASSO, as a regularization technique, provides competitive results, with R-squared scores of 0.819 and 0.821, and RMSE values of 40.774 and 0.462. So, the criteria confirmed that PR can be chosen the most accurate model for description of density and solubility.

Based on the outcomes of modeling, PR model was used as the model for generating the 3D Response Surfaces of two outputs, which are shown in Figs. 4 and 5. Also, the individual effect of inputs on both outputs visualized in Figs. 6, 7, 8 and 9. The results revealed that the solubility is increased with pressure of solvent as it behaves like gas solvents and its density varies with pressure unlike organic liquid solvents. This is indeed an important advantage of supercritical fluids whose solubility can be tuned with manipulating process pressure in addition to the temperature. On the other hand, the temperature reduces the solvent density (see Fig. 7) which has negative effect on the solubility, while the solubility is enhanced with increasing temperature which is due to the various phenomena involved in the solubility before and after cross-over pressure point in the system[30]. For determination of optimum point of operation, some economical evaluations are needed to find the cost of operation at each pressure and temperature.

Conclusion

In this research study, we compared three models - Polynomial Regression (PR), Extreme Gradient Boosting (XGB), and LASSO - for estimating SC-CO_2 density and niflumic acid solubility using temperature and pressure inputs. PR emerged as the most accurate model with R-squared scores of 0.992 for density and 0.969 for solubility, achieving low RMSE values of 12.203 and 0.256, respectively. XGB also performed well with R-squared scores of 0.927 and 0.930, and RMSE values of 28.623 and 0.286. LASSO demonstrated competitive results, with

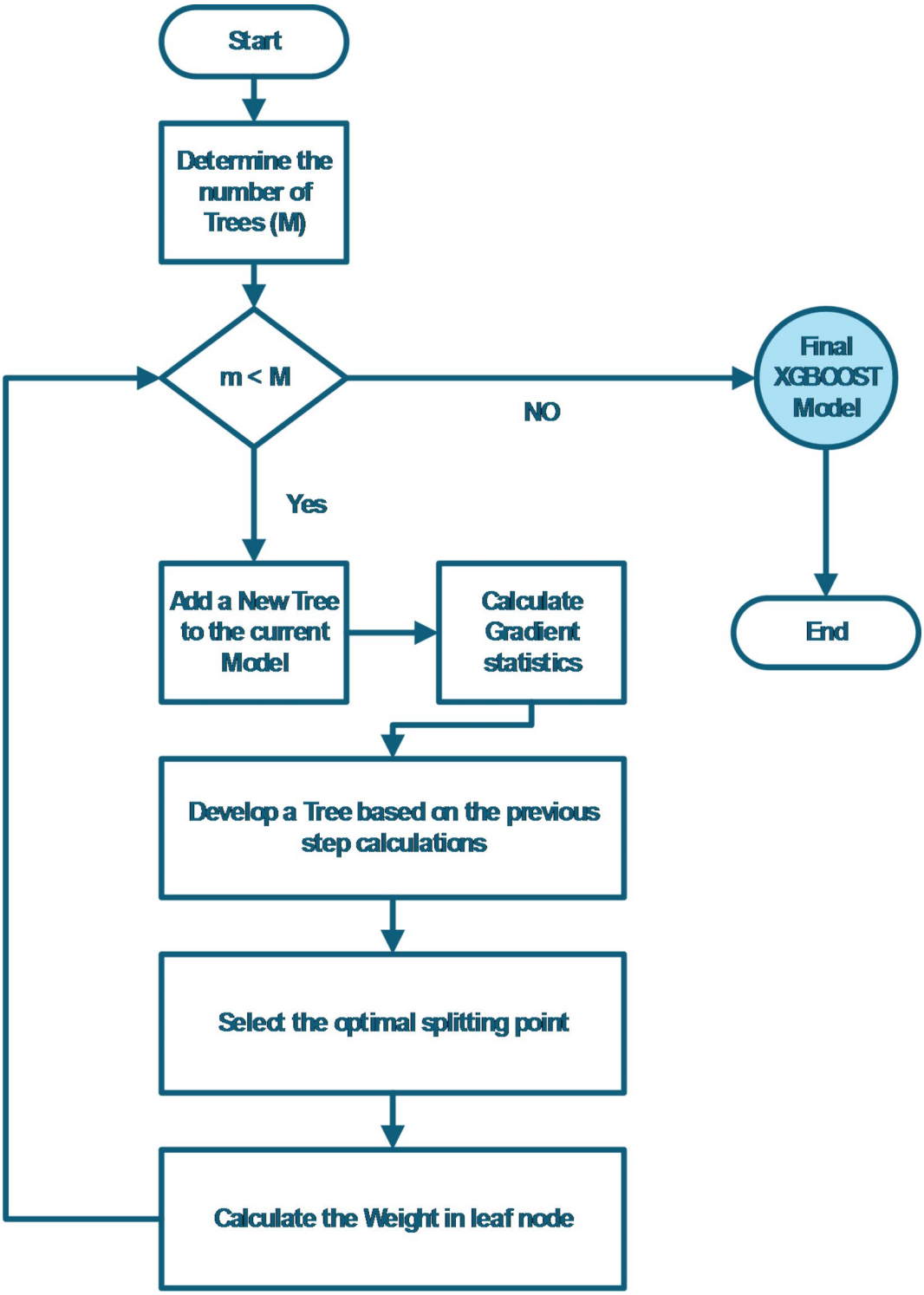

Fig. 1. The XGBoost Flowchart.

R-squared scores of 0.819 and 0.821, and RMSE values of 40.774 and 0.462. The BMO algorithm improved the models' performance. Overall, PR is the recommended model for accurate and interpretable predictions in these applications. The findings have practical implications for materials science, chemical engineering, and pharmaceutical research, supporting informed decision-making and process optimization. The developed methodology can be used as a generalized approach for data-driven decision making in pharmaceutical processing.

Model	*R*-squared Score	RMSE	Max Error
PR	0.992	12.203	20.803
XGB	0.927	28.623	60.604
LASSO	0.819	40.774	97.099

Table 1. SC-CO_2 density estimation by ML models.

Model	*R*-squared Score	RMSE	Max Error
PR	0.969	0.256	0.578
XGB	0.930	0.286	0.394
LASSO	0.821	0.462	1.037

Table 2. Solubility estimation by ML models.

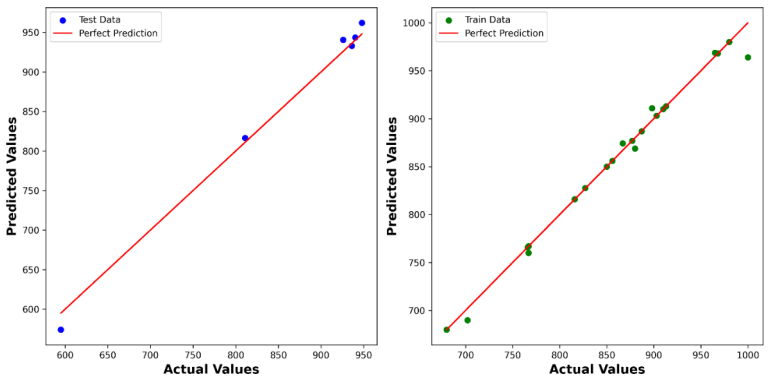

Fig. 2. Train and Test Results of Predicted and Actual values of Solvent Density using PR model.

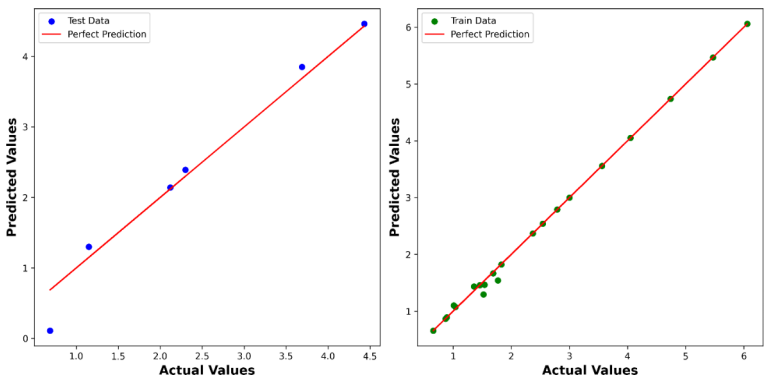

Fig. 3. Train and Test Results of Predicted and Actual values of Solubility using PR model.

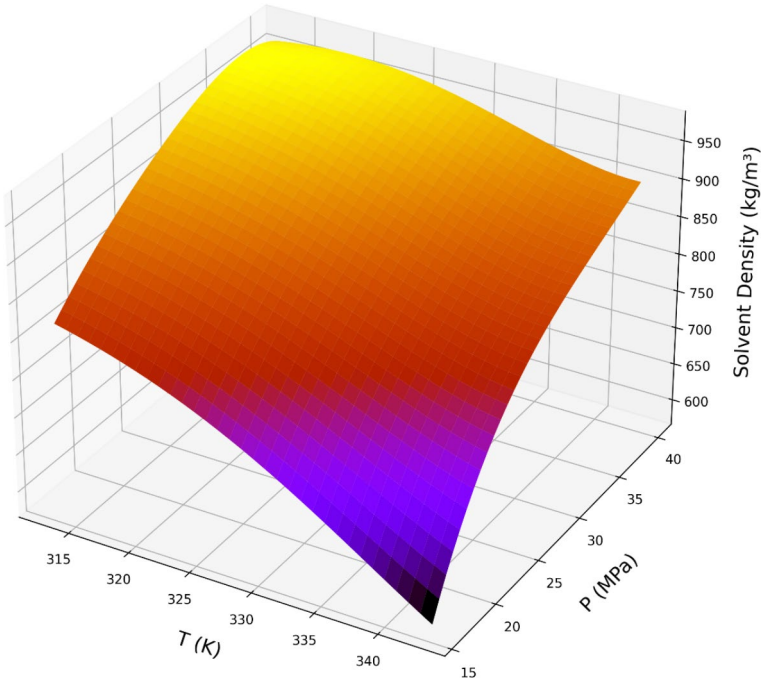

Fig. 4. Response Surface of Solvent Density generated using PR model.

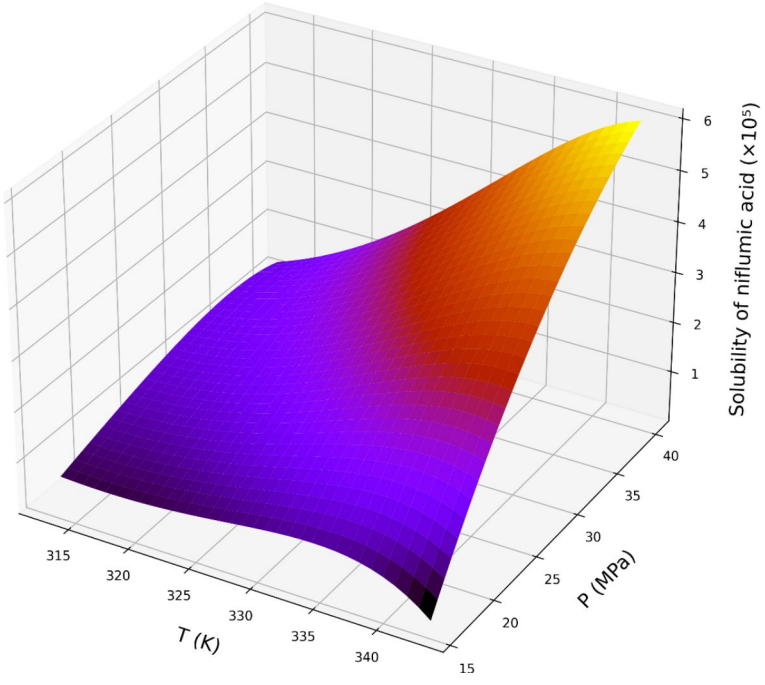

Fig. 5. Response Surface of niflumic acid solubility generated using PR model.

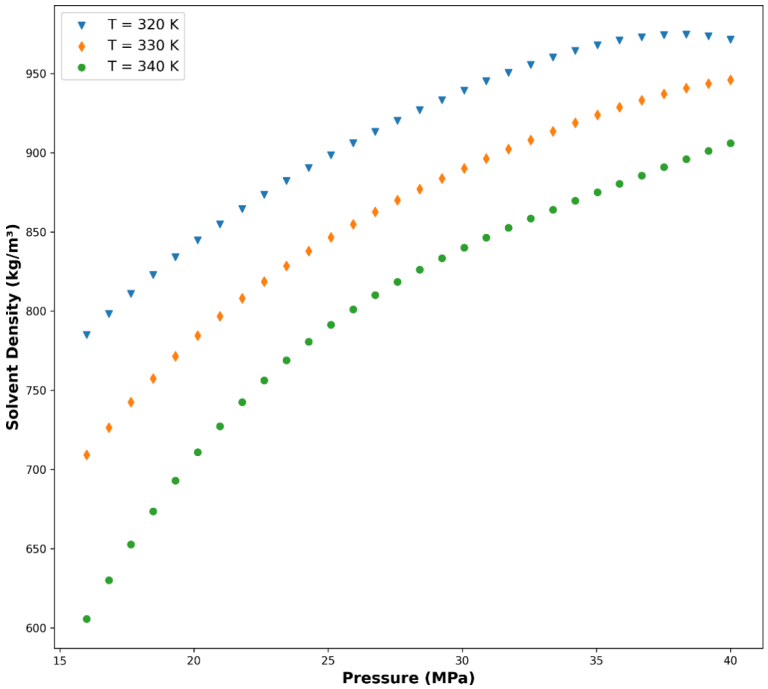

Fig. 6. Solvent Density based on Pressure keeping Temperature constant on different levels.

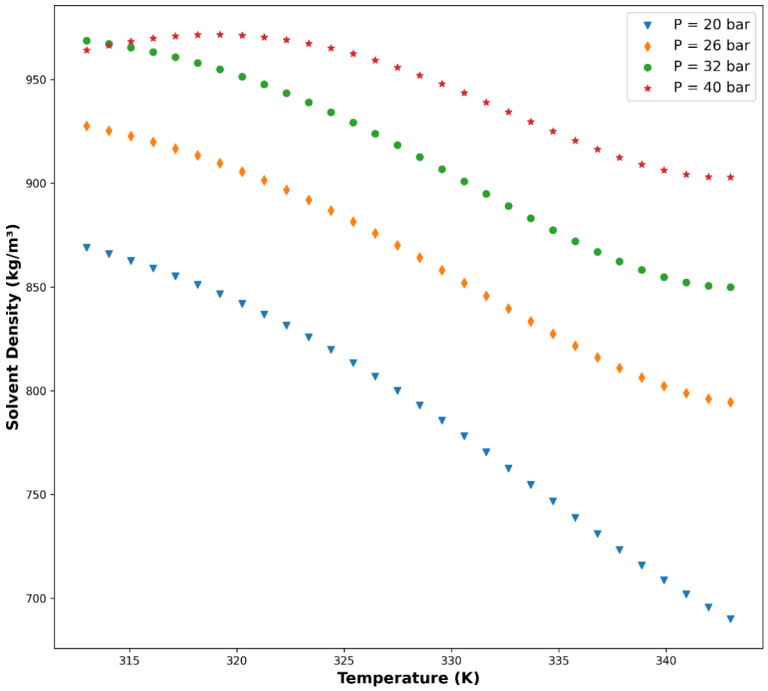

Fig. 7. Solvent Density based on Temperature keeping Pressure constant on different levels.

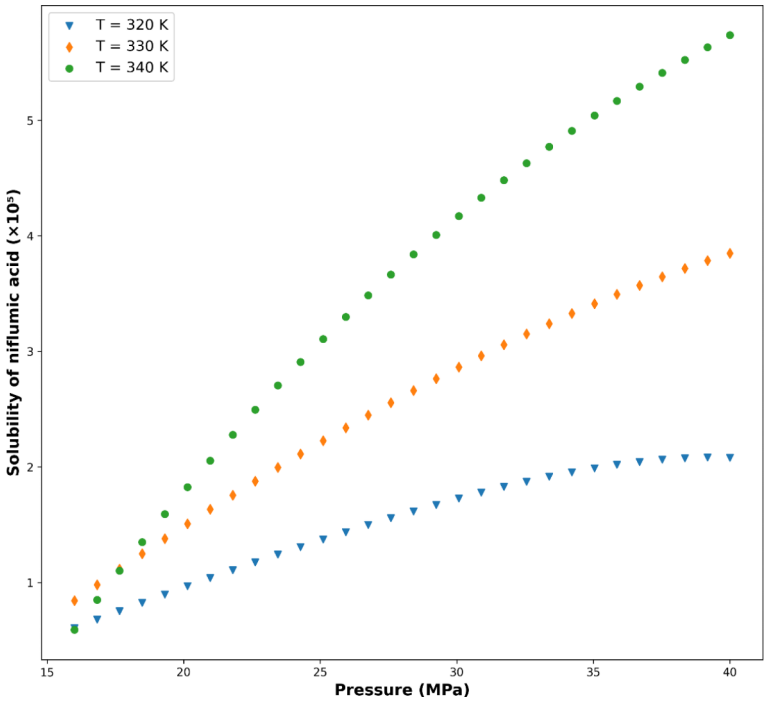

Fig. 8. Niflumic acid solubility based on Pressure keeping Temperature constant on different levels.

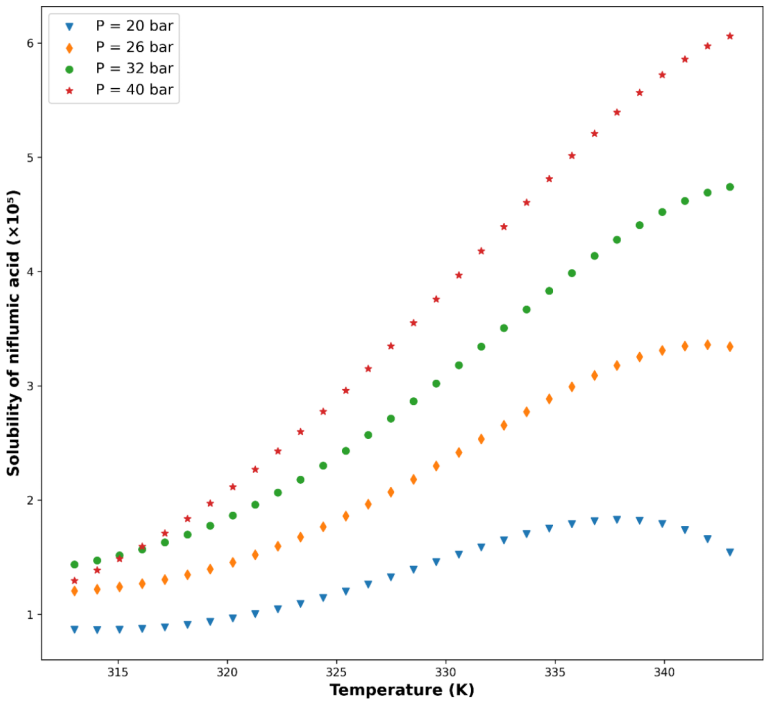

Fig. 9. Niflumic acid solubility based on Temperature keeping Pressure constant on different levels.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Received: 29 August 2024; Accepted: 26 September 2024
Published online: 01 October 2024

References

1. Darwich, M. et al. An approach for pH-independent release of poorly soluble ionizable drugs using hot-melt extrusion. *J. Drug Deliv. Sci. Technol.* **100**, 106027 (2024).
2. Wang, H. et al. Drug–drug co-amorphous systems: an emerging formulation strategy for poorly water-soluble drugs. *Drug Discovery Today*. **29** (2), 103883 (2024).
3. Wani, S. U. D. et al. Enhancing therapeutic potential of poor aqueous soluble herbal drugs through solid dispersion-An overview. *Phytomedicine Plus*. **1** (4), 100069 (2021).
4. Fan, W. et al. Application of the combination of ball-milling and hot-melt extrusion in the development of an amorphous solid dispersion of a poorly water-soluble drug with high melting point. *RSC Adv.* **9** (39), 22263–22273 (2019).
5. Bhatia, M. & Devi, S. Co-crystallization: a green approach for the solubility enhancement of poorly soluble drugs. *CrystEngComm*. **26** (3), 293–311 (2024).
6. Franco, P. & De Marco, I. Nanoparticles and nanocrystals by supercritical CO2-Assisted techniques for Pharmaceutical Applications: a review. *Appl. Sci.* **11** (4), 1476 (2021).
7. Sodeifian, G. et al. Determination of Gefitinib hydrochloride anti-cancer drug solubility in supercritical CO2: evaluation of sPC-SAFT EoS and semi-empirical models. *J. Taiwan Inst. Chem. Eng.* **161**, 105569 (2024).
8. Sodeifian, G. et al. Thermodynamic modeling and solubility assessment of oxycodone hydrochloride in supercritical CO2: semi-empirical, EoSs models and machine learning algorithms. *Case Stud. Therm. Eng.* **55**, 104146 (2024).
9. Abouzied, A. S. et al. Assessment of solid-dosage drug nanonization by theoretical advanced models: modeling of solubility variations using hybrid machine learning models. *Case Stud. Therm. Eng.* **47**, 103101 (2023).
10. An, F. et al. Machine learning model for prediction of drug solubility in supercritical solvent: modeling and experimental validation. *J. Mol. Liq.* **363**, 119901 (2022).
11. Chen, C. Artificial Intelligence aided pharmaceutical engineering: development of hybrid machine learning models for prediction of nanomedicine solubility in supercritical solvent. *J. Mol. Liq.* **397**, 124127 (2024).
12. Almehizia, A. A. et al. Numerical optimization of drug solubility inside the supercritical carbon dioxide system using different machine learning models. *J. Mol. Liq.* **392**, 123466 (2023).
13. Guan, S. & Wang, G. Drug discovery and development in the era of artificial intelligence: from machine learning to large language models. *Artif. Intell. Chem.* **2** (1), 100070 (2024).
14. Obaido, G. et al. Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects. *Mach. Learn. Appl.* **17**, 100576 (2024).
15. Muthukrishnan, R. & Rohini, R. *LASSO: A feature selection technique in predictive modeling for machine learning*. in *IEEE international conference on advances in computer applications (ICACA)*. 2016. IEEE. (2016).
16. Ranstam, J. & Cook, J. LASSO regression. *J. Br. Surg.* **105** (10), 1348–1348 (2018).
17. Ostertagová, E. Modelling using polynomial regression. *Procedia Eng.* **48**, 500–506 (2012).
18. Dutta, S. et al. *Robust counterfactual explanations for tree-based ensembles*. in *International Conference on Machine Learning*. PMLR. (2022).
19. Chen, T. et al. Xgboost: extreme gradient boosting. *R Package Version 0 4-2*. **1** (4), 1–4 (2015).
20. Banchero, M. & Manna, L. Solubility of fenamate drugs in supercritical carbon dioxide by using a semi-flow apparatus with a continuous solvent-washing step in the depressurization line. *J. Supercrit. Fluids*. **107**, 400–407 (2016).
21. Sulaiman, M. H. et al. Barnacles mating optimizer: a new bio-inspired algorithm for solving engineering optimization problems. *Eng. Appl. Artif. Intell.* **87**, 103330 (2020).
22. Alorf, A. A survey of recently developed metaheuristics and their comparative analysis. *Eng. Appl. Artif. Intell.* **117**, 105622 (2023).
23. Lee, J. H., Shi, Z. & Gao, Z. On LASSO for predictive regression. *J. Econ.* **229** (2), 322–349 (2022).
24. Ranciati, S., Roverato, A. & Luati, A. Fused graphical lasso for brain networks with symmetries. *J. Royal Stat. Soc. Ser. C: Appl. Stat.* **70** (5), 1299–1322 (2021).
25. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. (2016).
26. Guo, R. et al. Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. *Appl. Sci.* **10** (18), 6593 (2020).
27. Kutner, M. H. *Applied linear statistical models*. (2005).
28. Trevor, H., Robert, T. & Jerome, F. *The elements of statistical learning: data mining, inference, and prediction*. Spinger. (2009).
29. Seber, G. A. & Lee, A. J. *Polynomial regression*. Linear Regression Analysis, : pp. 165–185. (2003).
30. Alghazwani, Y. et al. Investigating the thermal enhancement of Levetiracetam solubility in the ternary system of supercritical carbon dioxide and ethanol. *J. Mol. Liq.* **411**, 125692 (2024).

Author contributions

Meixiuli Li: Writing, Investigation, Methodology, Conceptualization.Wenyan Jiang: Conceptualization, Formal analysis, Validation.Shuang Zhao: Resources, software, Visualization.Kai Huang: Writing, Investigation, Resources.Dongxiu Liu: Investigation, Writing, Formal analysis, Validation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.J.