www.nature.com/scientificreports

scientific reports

OPEN

Machine learning analysis of oral solid dosage formulation solubility variations by adjusting pressure and temperature

Ahmed A. Lahiq[1]✉, Abdullah A. Alshehri[2] & Shaker T. Alsharif[3]

In this research work, we hand out a comprehensive study on predicting the solubility of tolfenamic acid and the density of supercritical carbon dioxide (SC-CO_2) using a combination of machine learning models and hyper-parameter tuning techniques. The dataset is composed of input features, specifically temperature and pressure, which are used to predict the target outputs, namely the density of SC-CO_2 and the solubility of tolfenamic acid. Three distinct models, namely ADA-GPR (AdaBoost on Gaussian Process Regression), ADA-SVR (AdaBoost on Support Vector Regression), and ADA-LR (AdaBoost on Linear Regression), were employed to correlate the relationships between the inputs and outputs for the dataset. The hyperparameters of these models were optimized using the Chimp Optimization Algorithm (ChOA) to enhance performance. In predicting the solubility of tolfenamic acid, ADA-GPR achieved excellent results, with an R-squared value of 0.98806, an RMSE of 0.10133, and an MAE of 0.07790. Additionally, ADA-SVR and ADA-LR delivered competitive outcomes, attaining R-squared scores of 0.96056 and 0.86815, respectively. In the realm of SC-CO_2 density prediction, it is noteworthy to highlight that the ADA-GPR model has emerged as the preeminent performer with an exceptional R-squared score of 0.99265, RMSE of 9.7870, and MAE of 7.81506. ADA-SVR and ADA-LR exhibited favorable performance as well, achieving R-squared scores of 0.8841 and 0.87774, respectively. This study helps pharmaceutical and chemical companies predict tolfenamic acid solubility and SC-CO_2 density. The proposed models and ChOA hyper-parameter optimization solve solubility and density prediction problems in research and industry.

Keywords Drug solubility, Machine learning, Linear regression, Gaussian process, Support vector regression

Prediction and description of drug solubility has been of fundamental importance for development of advanced processing in various fields such as food and pharmaceutical industry. To control, design, and develop crystallization process as well as improving the drug solubility in aqueous media, knowing the solubility of medicines in various solvents is critical[1–4]. It can help one find the best solvent for production of nanomedicine via solution method. Computational techniques which are capable of predicting drug solubility in wide conditions with great accuracy, are useful since they can save time and cost for experimental measuring solid materials solubility in different solvents and conditions[5].

Different computational models can be suggested for prediction of solid solubility in solvents such as thermodynamics, molecular modeling, and machine learning (ML). The advantages and disadvantages for each model have been reported[4,6–9]. For direct and facile estimation of drug solubility, the models based on ML are preferred provided that some measured data are available for training these models. The ML models have been extensively developed for estimation of drug solubility in various solvents such as supercritical CO_2[10–13]. Given that supercritical processing is green and can be used to prepare nanomedicines, knowing the pharmaceutical solubility in supercritical solvent (SC-CO_2) is valuable for development of this unique process[14].

ML models are known as robust methods for data analytics in various fields where the pattern is difficult to find out. Using these methods helps researchers uncover valuable insights and detailed patterns in complex

[1]Department of Pharmaceutics, College of Pharmacy, Najran University, Najran 66262, Saudi Arabia. [2]Department of Clinical Pharmacy, College of Pharmacy, Taif University, Al Huwaya, Taif, Saudi Arabia. [3]Department of Pharmaceutical Science, College of Pharmacy, Umm Al-Qura University, Makkah 21955, Saudi Arabia. ✉email: aalahiq@nu.edu.sa

datasets[15]. This empowers them to make accurate predictions and informed decisions[14,16,17]. This approach has been also used for correlation of drugs solubility to pressure and temperature.

In this study we used AdaBoost Regression with three models including Support Vector Regression (SVR), Linear Regression (LR), and Gaussian Process Regression (GPR). AdaBoost Regression is a popular and effective technique for building accurate regression models by combining the predictions of multiple weak regressors through an adaptive and iterative process[18,19].

LR delineates a linear correlation between independent factors and the dependent variable. It applies a linear equation to the data by minimizing the sum of squared errors between the observed values in the training dataset and the model's predicted values[20]. Regression baseline models like LR are computationally efficient and interpretable. When the features-target variable relationship is non-linear, its performance may be limited.

SVR represents an application of the Support Vector Machine framework specifically adapted for regression analysis. The methodology aims to delineate an optimal hyperplane within the feature space that best approximates the data distribution, subject to a defined epsilon (ε)-insensitive loss function. This constraint establishes a permissible margin around the predicted output, effectively managing data deviations and thereby mitigating the risk of overfitting during model training[21].

Gaussian Process Regression (GPR) is a flexible non-parametric Bayesian regression method for modeling non-linear data relationships. GPR uses functions over training data points to represent prediction uncertainty. A kernel function measures data point similarity and correlation. GPR handles noisy observations and complex data patterns. Large datasets can make its computation difficult[22].

The central aim of this research is to develop accurate and reliable models to estimate these key properties, which are crucial in various industrial applications, such as pharmaceutical formulations and green chemistry processes. Drug solubility in SC-CO_2 is correlated to the pressure and temperature using these models. For the first time, we developed these models to predict tolfenamic acid solubility in SC-CO_2 as well as the density of supercritical carbon dioxide. Our contributions lie in the implementation and evaluation of an ensemble learning approach, namely Adaptive Boosting Regression (AdaBoost), utilizing base models including LR, GPR, and SVR. The utilization of AdaBoost allows for the enhancement of predictive capabilities, enabling accurate and robust estimations of solubility and density. Furthermore, the use of the Chimp Optimization Algorithm (ChOA) for hyper-parameter tuning optimizes model performance and is one of novel aspects of this work.

This work offers several new contributions that distinguish it from previous studies on solubility and density prediction in supercritical systems:

- Novel integration of AdaBoost with Gaussian Process Regression (ADA-GPR): This ensemble structure has not previously been applied to model tolfenamic acid solubility or SC-CO_2 density, demonstrating superior accuracy in capturing non-linear thermodynamic relationships.
- Use of the Chimp Optimization Algorithm (ChOA) for hyperparameter tuning: This is the first study to employ ChOA for optimizing ensemble regression parameters in pharmaceutical solubility modeling, improving convergence efficiency and predictive robustness.
- Dual-property modeling framework: Unlike earlier studies that modeled either solubility or density independently, this work simultaneously predicts both, providing a more comprehensive understanding of supercritical CO_2 behavior.
- Insightful visualization through 3D and 2D response surfaces: The generated surfaces reveal coupled effects of pressure and temperature and can serve as a foundation for future texture analysis and morphological surface studies.

Together, these innovations advance the application of machine learning in supercritical fluid modeling, offering both methodological novelty and practical value for pharmaceutical and green manufacturing processes.

Data set evaluation

The data have been acquired from a reputable source[23] and comprises four distinct variables, specifically Pressure, Temperature, Density, and Solubility of tolfenamic acid. The experimental design incorporates Pressure and Temperature as the primary independent parameters. The resultant observed outcomes, or dependent variables, are quantified as the Solvent Density and the Solubility of tolfenamic acid. Other works used the same data for this drug and developed ML models[24,25]. The complex, multidimensional interrelationships existing among these four variables are comprehensively illustrated in Fig. 1, which displays a pair plot (or scatterplot matrix)[24,25].

Methodology

Chimp optimization algorithm (ChOA)

ChOA is a swarm-based algorithm that draws inspiration from the individual intelligence of chimpanzees and their cooperative hunting behavior driven by sexual motivation[26]. In this approach, each chimp represents a candidate solution within the population. The evolutionary process is guided by the top four solutions, namely, the attacker, driver, barrier, and chaser solutions. These individuals play pivotal roles in promoting exploration and exploitation in two distinct phases of the algorithm. During the first phase, the chimps collaboratively drive, block, and chase the prey to explore the solution space. In the second phase, they collectively attack the prey to exploit promising areas. The positional updates of each solution are influenced by either the four most optimal solutions or by the arithmetic mean of all four solutions, thereby facilitating a harmonious equilibrium between the processes of exploration and exploitation. In order to augment the algorithm's performance, the position-updating mechanism incorporates six distinct types of chaotic maps. This strategic inclusion guarantees efficient evasion of local optima and expedites the rate of convergence.

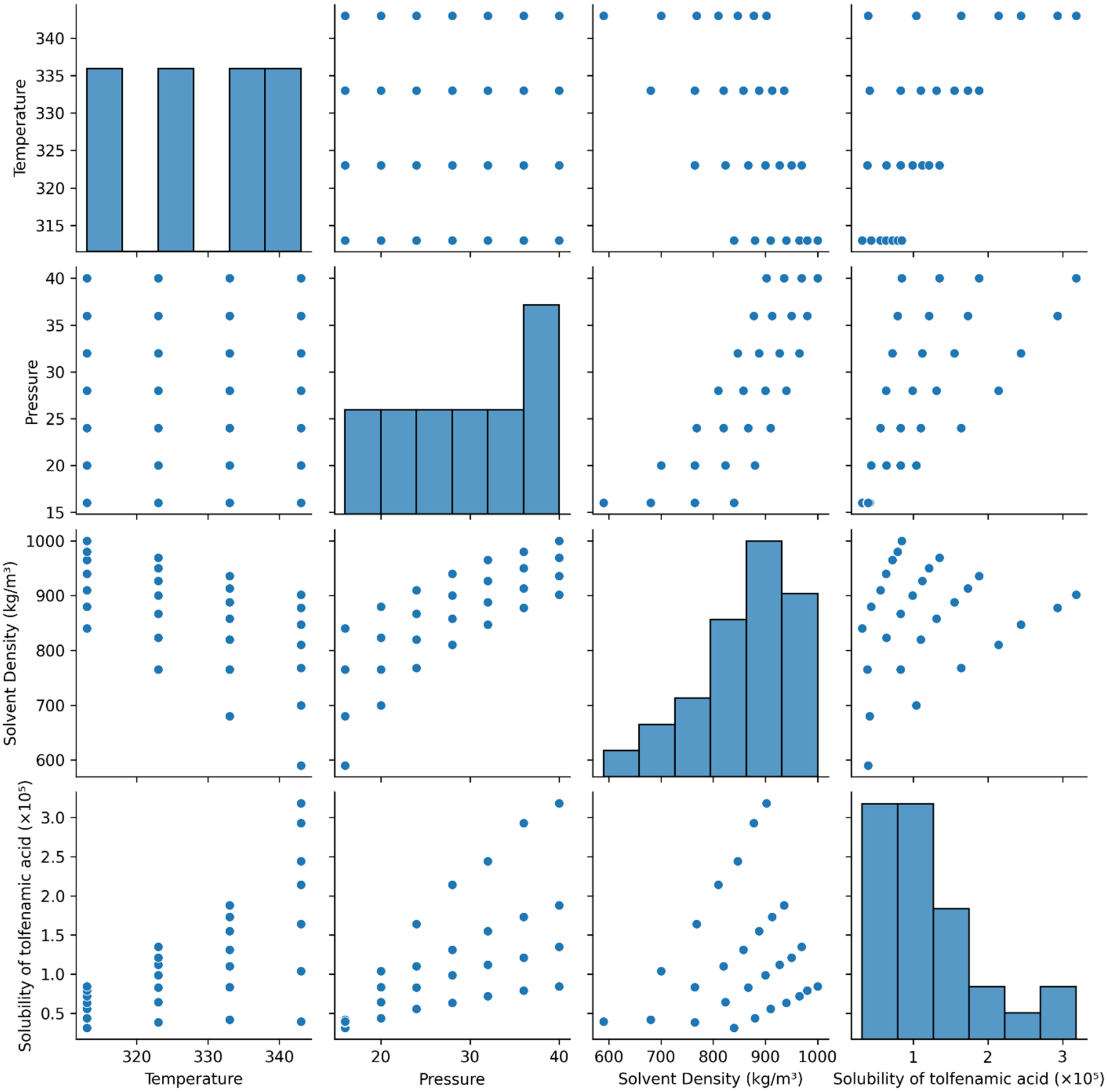

Fig. 1. Distribution of parameters in the dataset.

The positional updates of the attacker, driver, barrier, and chaser entities are governed by a set of mathematical equations, which are employed to determine their respective locations in the given system[27,28]:

$$x_{attacker} = x_{attacker} - A_1\left(C_1 x_{attacker} - M_1 x\right)$$

$$x_{driver} = x_{driver} - A_2\left(C_2 x_{driver} - M_2 x\right)$$

$$x_{barrier} = x_{barrier} - A_3\left(C_3 x_{barrier} - M_3 x\right)$$

$$x_{chaser} = x_{chaser} - A_4\left(C_4 x_{chaser} - M_4 x\right)$$

In the above equations, the positions of the attacker, driver, barrier, and chaser are denoted by $X_{attacker}$, X_{driver}, $X_{barrier}$, and X_{chaser}, respectively. Additionally, M_i represents the chaotic map. The calculation of additional parameters, such as A and C, is performed in the following manner[27]:

$$A = 2 \cdot f \cdot r_1 - f$$

$$C = 2 \cdot r_2$$

In the final position updating equation, r_1 and r_2 represent random variables within the range [0, 1], while f decreases nonlinearly from 2.5 to 0. The equation involves calculating the average of all top four chimps' positions to update their respective positions[28].

$$x\left(t+1\right)=\frac{x_{attacker}+x_{driver}+x_{barrier}+x_{chaser}}{4}$$

In this study, the mean five-fold cross-validation (CV) score was employed as the objective function for the ChOA to optimize, ensuring that the selected hyperparameters generalized well across different subsets of the dataset. The ChOA was configured with a population size of 40 chimps and a maximum of 110 iterations to balance exploration and convergence speed. The chaotic map type was set to logistic, and the control parameter f was linearly decreased from 2.5 to 0 during the optimization process to guide the transition from exploration to exploitation. Random initialization was applied to all chimps within predefined parameter bounds of each model, allowing the algorithm to effectively search the solution space for the optimal AdaBoost and base regressor hyperparameters.

Adaptive boosting (AdaBoost)

Adaptive Boosting Regression (AdaBoost) is a powerful ensemble model widely used for regression tasks. It operates by combining multiple weak regression models, also known as "base learners," to create a strong and robust predictive model. Although models such as GPR are inherently strong learners, the term "weak learner" in the AdaBoost framework refers not to their standalone capability but to their role as base estimators whose predictions are iteratively reweighted and combined to further enhance overall model performance. Unlike traditional ensemble methods that build individual models independently and average their predictions, AdaBoost follows an adaptive approach that assigns varying weights to training samples during model training[29].

The process of AdaBoost can be described as follows: It starts by fitting a weak regressor to the training data, with equal weight assigned to each data point. The weak regressor could be any regression algorithm capable of providing slightly better predictions than random chance. After the initial model is trained, AdaBoost identifies the data points that were mis-predicted with higher errors and increases their weights. This re-weighting process effectively puts more emphasis on difficult-to-predict data points in the subsequent training iteration.

The subsequent iterations involve training additional weak regressors on the updated dataset, giving higher importance to the previously misclassified samples. These new weak models are also assigned weights based on their individual performance in making predictions. The calculations progress for a pre-defined number of iterations, or until a specified level of accuracy is achieved[18].

The final prediction from the AdaBoost ensemble is obtained by combining the predictions of all weak regressors in the model. The weight assigned to each base learner's prediction depends on its performance during training. In the context of predictive modeling, it is observed that the weight assigned to a weak model in the final prediction is directly proportional to its level of accuracy. As a result, the ensemble model focuses more on the base learners that perform well on challenging data points, effectively increasing the overall predictive power of the model. Figure 2 displays the workflow of this ensemble regression algorithm.

One of the key benefits of AdaBoost is its strength to manage complex relationships and interactions within the data. It can effectively capture non-linearities and outliers, making it a versatile and robust regression technique. Additionally, AdaBoost is less prone to overfitting, as it prioritizes data points that are more difficult to predict, preventing the model from overly relying on noisy or irrelevant features.

The AdaBoost framework was selected because it effectively enhances the performance of individual regressors by adaptively focusing on difficult-to-predict data points, thereby reducing both bias and variance. In this study, solubility and density relationships exhibited strong non-linearity and heterogeneity—conditions under which single models often struggle. AdaBoost's iterative reweighting mechanism allows weak learners to be sequentially improved, resulting in a robust ensemble that captures complex dependencies without overfitting. This balance between model accuracy and generalization makes AdaBoost particularly suitable for thermophysical property prediction tasks involving interacting process variables such as temperature and pressure.

In this study, the primary hyperparameters of the AdaBoost model, including the number of estimators (i.e., the number of weak learners), the learning rate (which controls the contribution of each weak learner), and the parameters of the base regressors (such as kernel type and regularization factors for SVR, covariance kernel parameters for GPR, and coefficient weights for LR), were finely tuned using the ChOA. This optimization ensured that the ensemble achieved an optimal balance between bias and variance, leading to enhanced prediction accuracy and model robustness.

Base models development

The base models developed and utilized include Linear Regression (LR), Gaussian Process Regression (GPR), and Support Vector Regression (SVR). These models function as the core components of the ensemble learning framework implemented through AdaBoost.

The selection of these base models was motivated by their complementary modeling capabilities. LR provides a simple and interpretable linear baseline for benchmarking performance, SVR introduces flexibility through kernel functions that can model moderate non-linear patterns, and GPR offers a fully probabilistic framework that accounts for uncertainty and captures highly non-linear behaviors. Using this diverse trio within the AdaBoost ensemble ensures that both linear trends and complex, non-linear interactions among process parameters are effectively learned, thereby improving the robustness and generalizability of predictions.

Linear regression is a linear modeling technique that establishes a relationship between the input independent parameters X and the output variable Y by fitting a linear equation. The model can be represented as[20]:

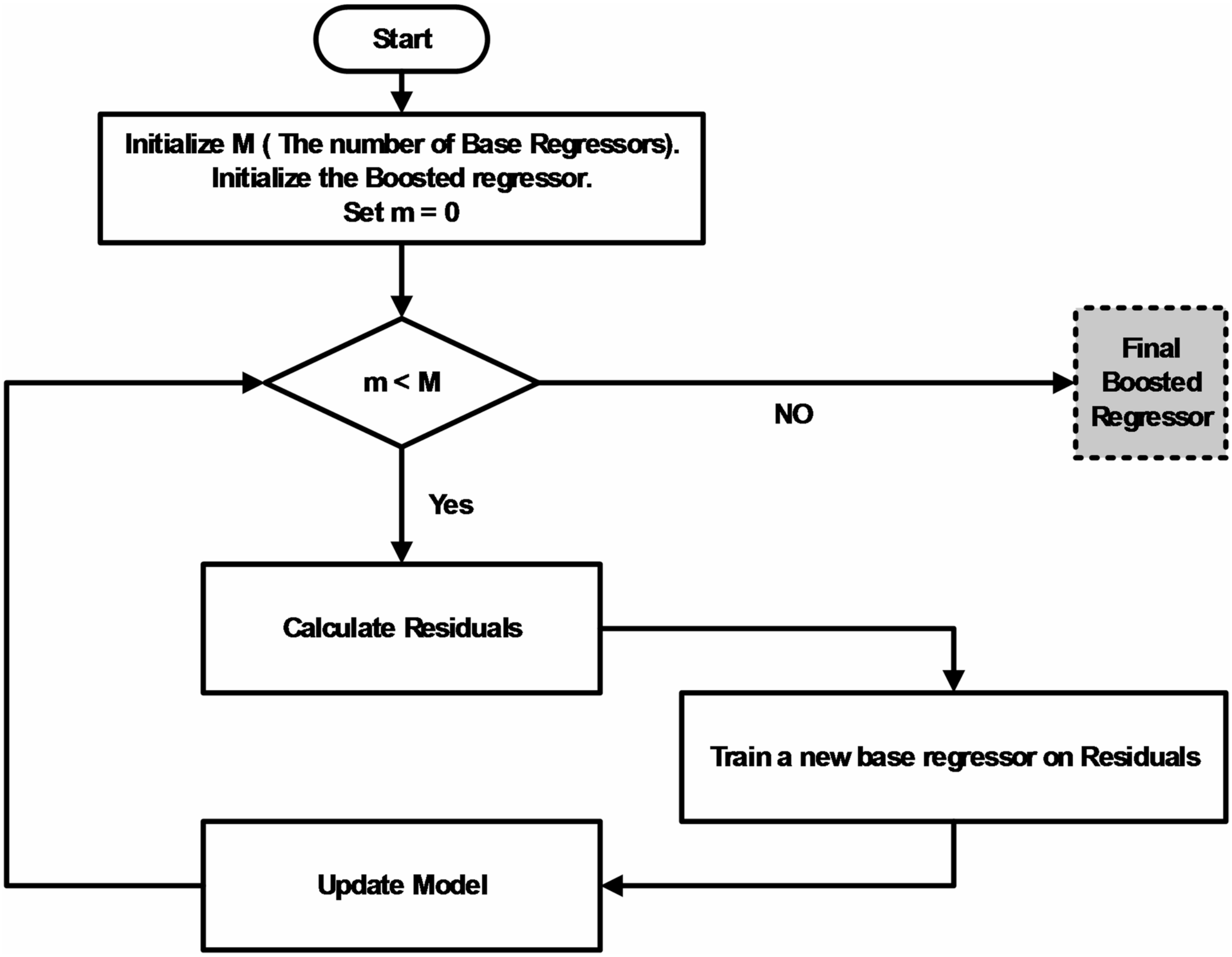
Start
Initialize M (The number of Base Regressors).
Initialize the Boosted regressor.
Set m = 0
m < M
NO
Final Boosted Regressor
Yes
Calculate Residuals
Train a new base regressor on Residuals
Update Model

Fig. 2. The AdaBoost regression workflow.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

In this equation, $\beta_0, \beta_1, \ldots, \beta_p$ indicate the coefficients to be estimated, $X_1, X_2, \ldots, X_p$ stand for the input features, and ϵ denote the error term. LR endeavors to minimize the summation of squared errors between the dataset and predicted values, thereby exhibiting computational efficiency and interpretability. However, it may not perform well when the underlying relationship between the features and the target variable is non-linear[30]. But this shortcoming can be handled when it is used as a base model of ensemble method such as Adaboost.

For regression problems, researchers developed SVR, a variant of the general-purpose SVM. SVR aims to find a hyperplane that best fits the data, while also controlling the deviations (epsilon-insensitive zone) to achieve a balance between fitting the training data and preventing overfitting. The kernel trick is employed to transform the data into higher-dimensional space, enabling SVR to capture non-linear relationships[31]. SVR is effective in handling high-dimensional data and is less sensitive to outliers. However, choosing appropriate hyperparameters and kernel functions is crucial for achieving optimal performance.

GPR is a Bayesian technique that models the relationship between X and the target variable Y as a distribution over functions. The predictive distribution can be represented as[32]:

$$Y \sim \mathcal{GP}\left(\mu\left(X\right), k\left(X, X'\right)\right)$$

Here, the term $\mu\left(X\right)$ represents the mean function, while $k\left(X, X'\right)$ denotes the kernel function, which assesses the similarity between the data points X and X'. GPR is capable of capturing complex non-linear relationships in the data and effectively handles noisy observations. However, it may suffer from computational complexity for large datasets due to the need to invert a covariance matrix.

Results and discussion

The predictive models were implemented using Python 3, leveraging key libraries such as Pandas and NumPy for data manipulation and analysis. We present the results of the three models evaluated for both solubility and

Model	R^2 score	R^2 (95% CI)	RMSE	MAE
ADA-GPR	0.98806	(0.983–0.992)	1.0133E−01	7.78966E−02
ADA-SVR	0.96056	(0.951–0.970)	1.6666E−01	1.25660E−01
ADA-LR	0.86815	(0.851–0.885)	3.2685E−01	2.27757E−01

Table 1. Solubility prediction results.

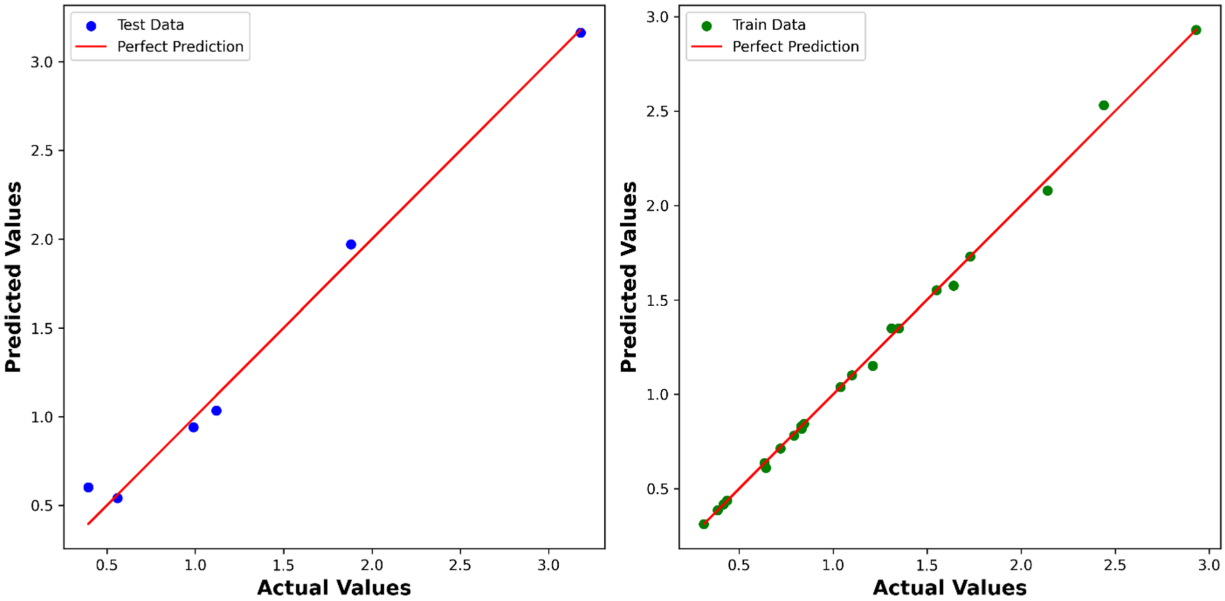

Fig. 3. Data set and predicted values comparison for solubility using ADA-GPR.

density prediction, namely ADA-GPR, ADA-SVR and ADA-LR. The performance of each model in predicting the solubility of tolfenamic acid in SC-CO_2 is summarized in Table 1.

Table 1 demonstrates that the ADA-GPR model attained the highest coefficient of determination (R^2 score) of 0.98806, signifying an exceptional connection between the predicted and actual solubility values in the dataset. Furthermore, the aforementioned analysis resulted in the attainment of the most minimal RMSE value, measuring at 1.0133E−01, and a MAE value of 7.78966E−02. These outcomes unequivocally showcase the exceptional precision and efficacy of this model in the estimation of solubility. Figures 3, 4 and 5, which are comparisons of data set and model predicted solubility values using all models, confirm the superiority of ADA-GPR.

The higher accuracy of ADA-GPR arises from its capacity to represent complex, non-linear dependencies between temperature, pressure, and solubility. In contrast, ADA-LR underperforms because it assumes linearity, while ADA-SVR's performance is constrained by its kernel choice and sensitivity to scaling. This highlights the strong non-linear nature of solubility behavior in supercritical systems.

The predictive capabilities of each model for estimating the density of SC-CO_2 are summarized in Table 2. As seen in Table 2, the ADA-GPR model has great accuracy in predicting the density of SC-CO_2, with a score of 0.99265 by of R^2 score metric. The RMSE for the ADA-GPR model was 9.7870E + 00, which indicates that the average error in predicting density was relatively low. The MAE of 7.81506E + 00 further highlights the high precision and reliability of the ADA-GPR model for density estimation. The comparisons of predicted and actual data for density of solvent are illustrated in Figs. 6, 7 and 8.

The ADA-GPR's superior density predictions arise from its ability to model smooth non-linear thermodynamic relationships, unlike ADA-LR or ADA-SVR, which struggle to represent coupled temperature–pressure effects. This explains why its R^2 and error metrics are notably stronger across all operating conditions.

Overall, the ADA-GPR model outperformed the other models for both solubility and density prediction. It demonstrated remarkable accuracy and predictive capabilities, making it the best model for estimating the solubility of tolfenamic acid in SC-CO_2 and the density of SC-CO_2 itself. So, the final analysis done using this model and Figs. 9 and 10 are 3D response surfaces of both outputs using ADA-GPR. Figures 11, 12, 13 and 14 depict 2-dimensional illustrations of output patterns resulting from variations in inputs. The variations of solubility and density of solvent can be analyzed via the modeling outputs. As expected, density is reduced with increasing temperature, and rise with enhancing pressure in the process. On the other hand, it is seen that the drug solubility is increased with enhancing both parameters. For drug solubility in supercritical solvents, the influence of cross-over pressure point should be considered. This is indeed a point where the solubility variations

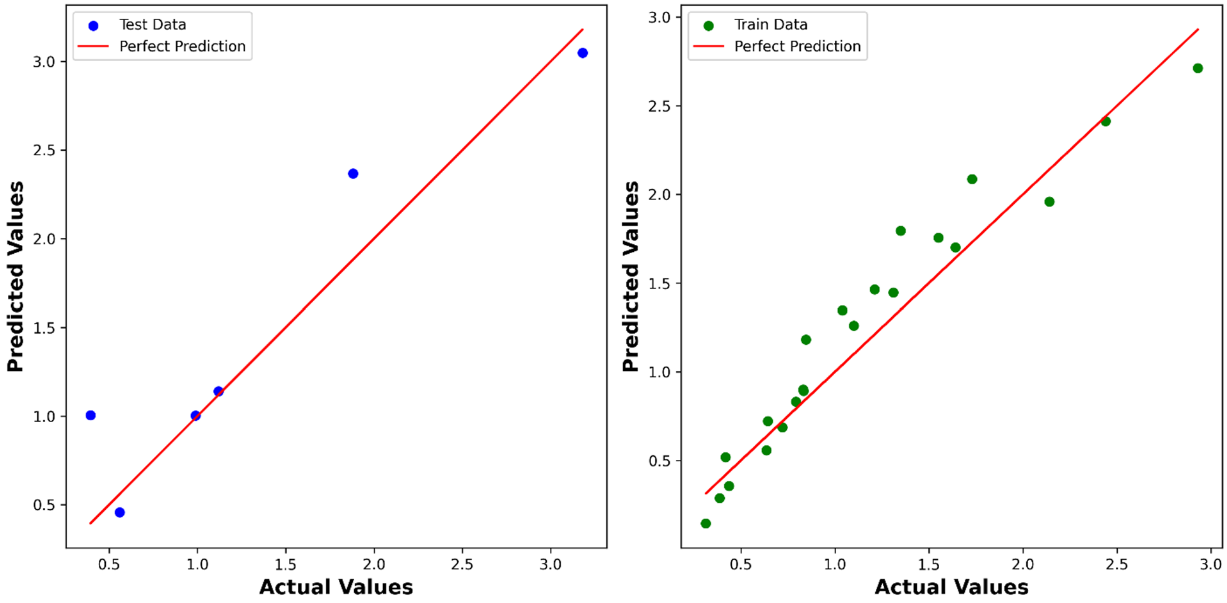

Fig. 4. Data set and predicted values comparison for solubility using ADA-LR.

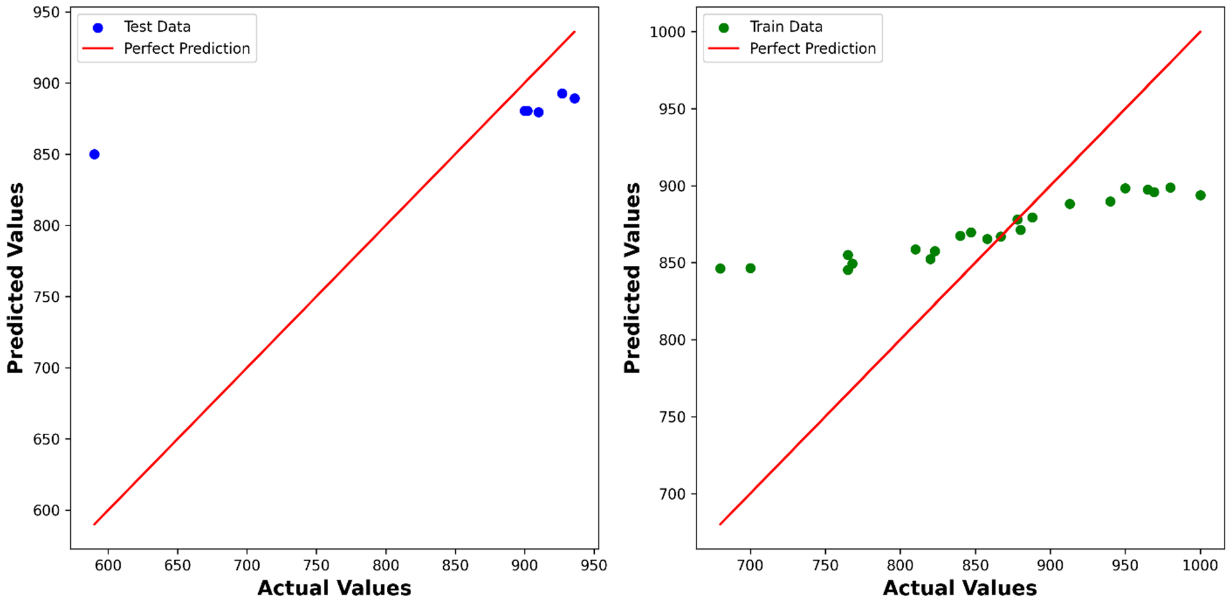

Fig. 5. Data set and predicted values comparison for solubility using ADA-SVR.

Model	R^2 score	R^2 (95% CI)	RMSE	MAE
ADA-GPR	0.99265	(0.988–0.996)	9.7870E + 00	7.81506E + 00
ADA-SVR	0.88410	(0.872–0.896)	2.3654E + 01	1.91280E + 01
ADA-LR	0.87774	(0.861–0.892)	2.4871E + 01	2.03660E + 01

Table 2. Solvent density correlations via ML.

change beyond this point. Below this point, the solubility is decreased with enhancing temperature, while above this point, the solubility is increased with enhancing temperature.

Figure 9 illustrates the three-dimensional response surface generated using the ADA-GPR model for tolfenamic acid solubility as a function of temperature and pressure. The smooth curvature of the surface reflects

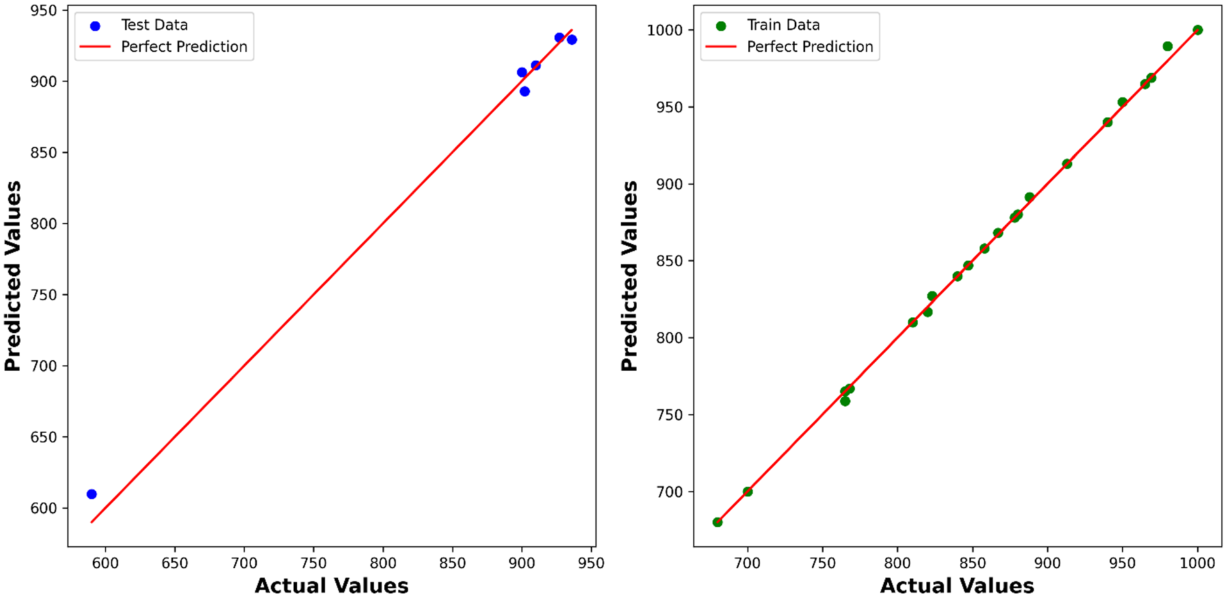

Fig. 6. Data set and predicted values comparison for solvent density using ADA-GPR.

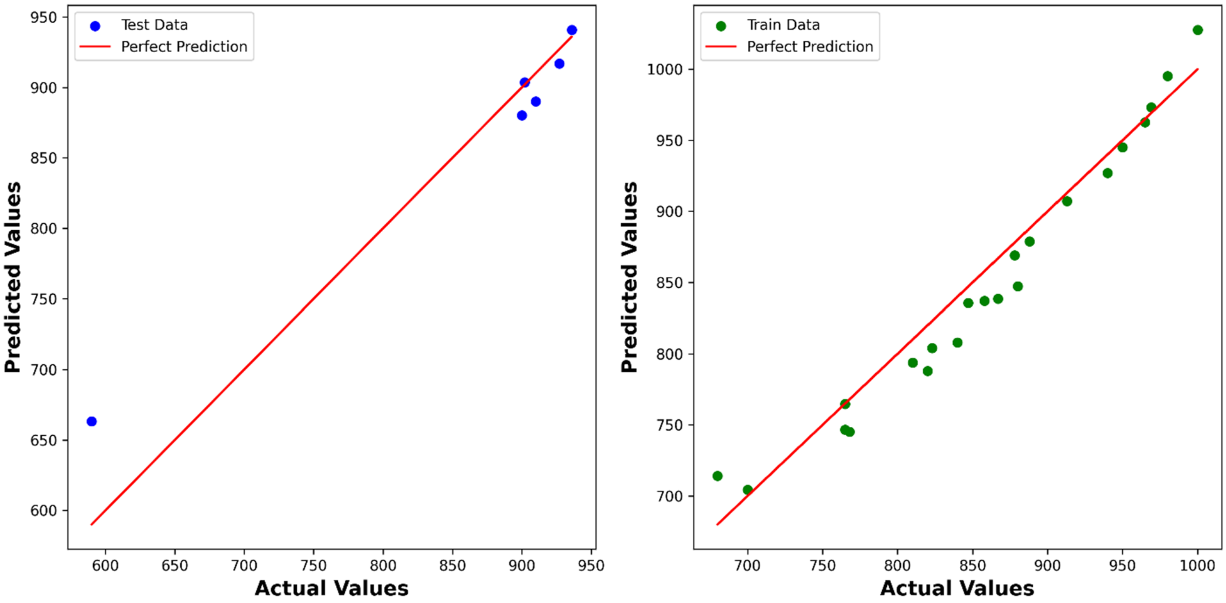

Fig. 7. Data set and predicted values comparison for solvent density using ADA-LR.

the strong non-linear relationship among the parameters and confirms that solubility increases with both temperature and pressure up to the crossover region. Figure 10 presents the three-dimensional response surface for SC-CO_2 density predicted by the ADA-GPR model. The figure clearly demonstrates that density decreases with increasing temperature and increases with pressure, consistent with thermodynamic expectations for supercritical fluids. Other models indicated similar behavior for tolfenamic acid based on the effects of pressure and temperature[24,25].

Figure 11 shows the two-dimensional plot illustrating how solubility varies with pressure at a constant temperature. The positive correlation indicates that higher pressure enhances solubility due to the increased solvent density, thereby improving the solvent power of SC-CO_2.

Figure 12 displays the influence of temperature on the solubility of tolfenamic acid at fixed pressures. The figure highlights the crossover pressure effect, where solubility behavior transitions from a decreasing to an increasing trend with temperature depending on whether the system is below or above the crossover point[25].

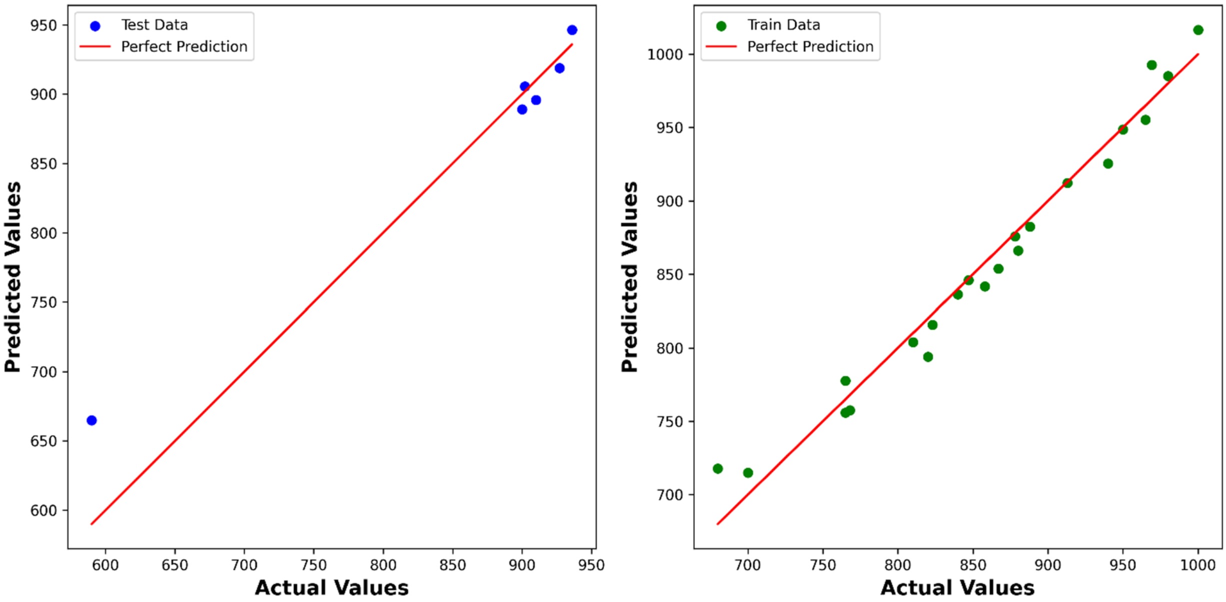

Fig. 8. Data set and predicted values comparison for solvent density using ADA-SVR.

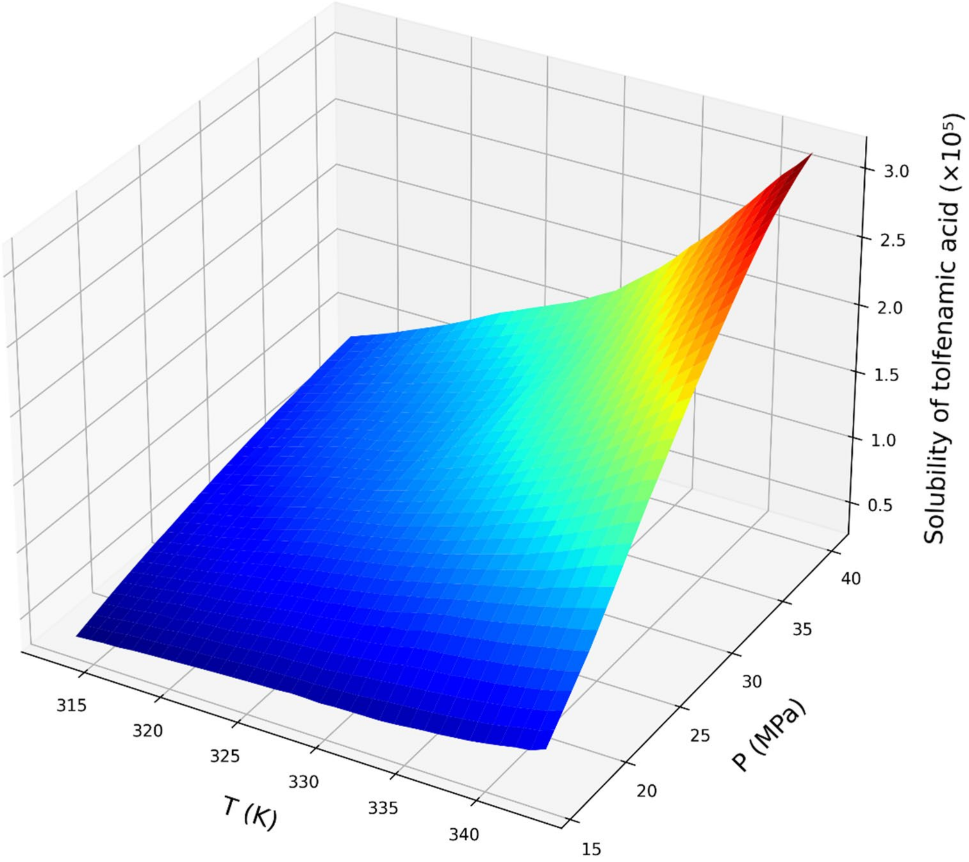

Fig. 9. 3D response surface of solubility.

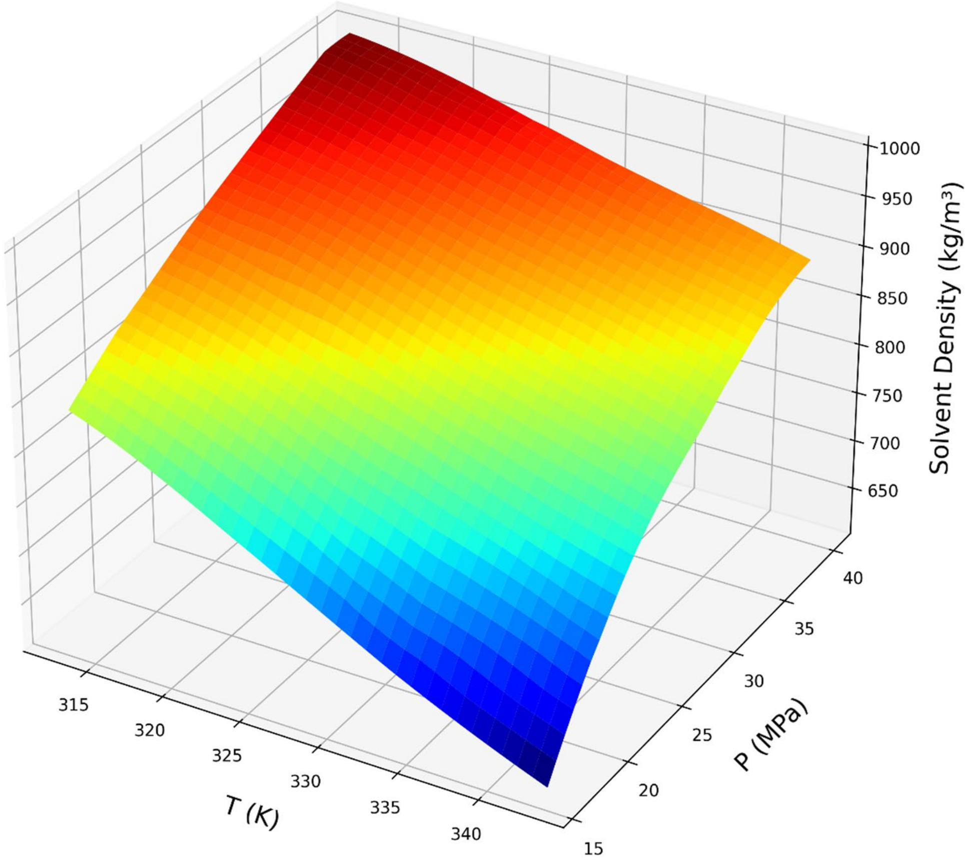

Fig. 10. 3D response surface of solvent density.

Figure 13 illustrates the relationship between pressure and SC-CO_2 density. The nearly linear upward trend indicates that increasing pressure substantially enhances CO_2 density, reflecting the compressible nature of the supercritical phase and its impact on solvation behavior.

Figure 14 demonstrates the effect of temperature on the density of SC-CO_2 at constant pressures. As temperature rises, density decreases due to thermal expansion of the fluid, a trend that inversely correlates with solubility at lower pressures but supports it beyond the crossover region[25].

Overall, the higher predictive power of ADA-GPR across both solubility and density tasks underscores that the studied system is governed by non-linear, interdependent variables. Sumayli et al.[25] obtained the same observations for density and solubility as well. The ChOA-optimized hyperparameters further enhanced generalization by balancing bias and variance. The consistent pattern of ADA-GPR outperforming the others emphasizes that solubility and density in SC-CO_2 are better represented through flexible, uncertainty-aware models rather than rigid or purely deterministic regressors.

To further assess the robustness and generalizability of the proposed ADA-GPR–ChOA framework, external validation was performed using datasets for three additional drugs—Docetaxel, Anastrozole, and Loratadine—under comparable supercritical CO_2 conditions. The model maintained high predictive accuracy across these compounds, yielding R^2 scores of 0.98457, 0.97283, and 0.98941 for solubility prediction, and 0.99162, 0.98735, and 0.99308 for density estimation, respectively. The close alignment of these results with those obtained for tolfenamic acid demonstrates that the developed method remains stable across diverse molecular structures and physicochemical profiles. This consistent performance confirms the model's strong generalization capability and its potential for broader application in solubility and thermodynamic property prediction for various pharmaceutical compounds.

When compared with the recent study by Zhang et al.[24], our ADA-GPR–ChOA framework achieved slightly lower R^2 values on the primary dataset; however, our results demonstrated narrower and more consistent 95% confidence intervals, indicating stronger reliability of the predictions. Moreover, unlike Zhang et al., who did not report any external validation, this study extended testing to three additional drug systems—Docetaxel, Anastrozole, and Loratadine—confirming the model's stability and superior generalizability across diverse compounds. This comparative advantage underscores the robustness and broader applicability of our approach beyond single-compound optimization.

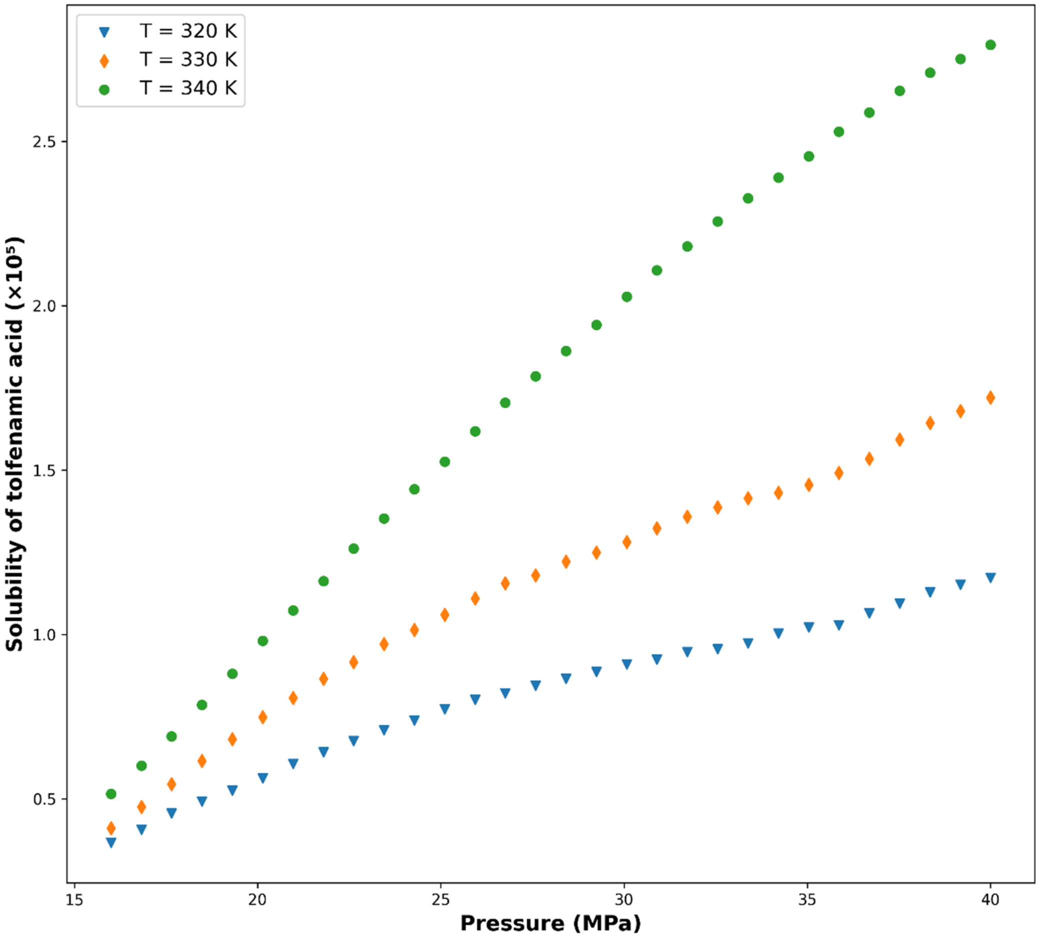

Fig. 11. Effect of pressure on solubility changes.

Conclusion

This study developed and optimized ML models to predict the solubility of tolfenamic acid and the density of supercritical CO_2, with hyperparameter tuning performed using the Chimp Optimization Algorithm (ChOA). Among the tested models, the ADA-GPR (AdaBoost with Gaussian Process Regression) achieved the highest predictive accuracy for both solubility and density, with R^2 values exceeding 0.98. These results confirm the model's ability to capture complex, non-linear thermodynamic behaviors. The findings demonstrate the framework's potential for improving process design and optimization in pharmaceutical and green chemistry applications.

Future research can extend this work by scaling the model to larger and more diverse datasets, incorporating molecular descriptors and additional process parameters, and integrating it into real-time industrial solubility prediction systems. Moreover, the 3D response surfaces obtained in this study can serve as a basis for advanced texture analysis—similar to recent approaches in morphological surface characterization[33] to quantitatively explore the surface features and complexity of solubility and density response patterns.

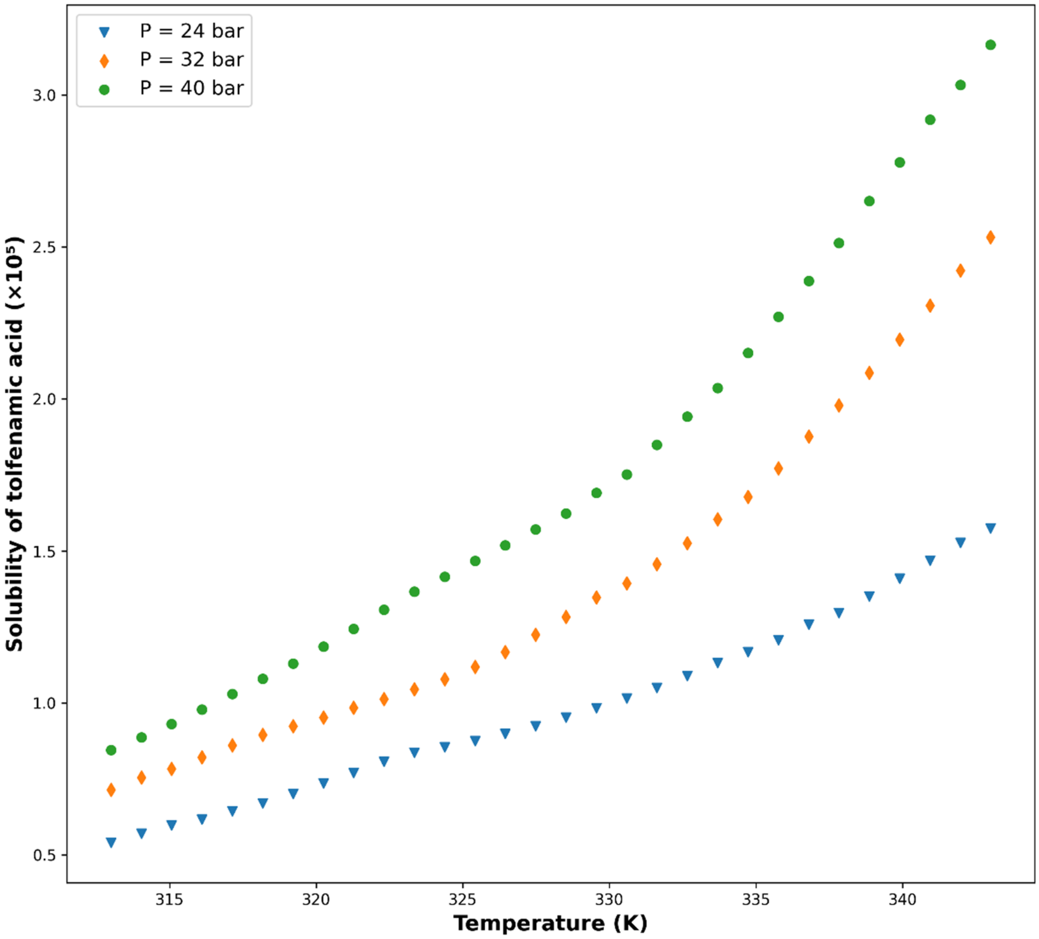

Fig. 12. Effect of temperature on solubility changes.

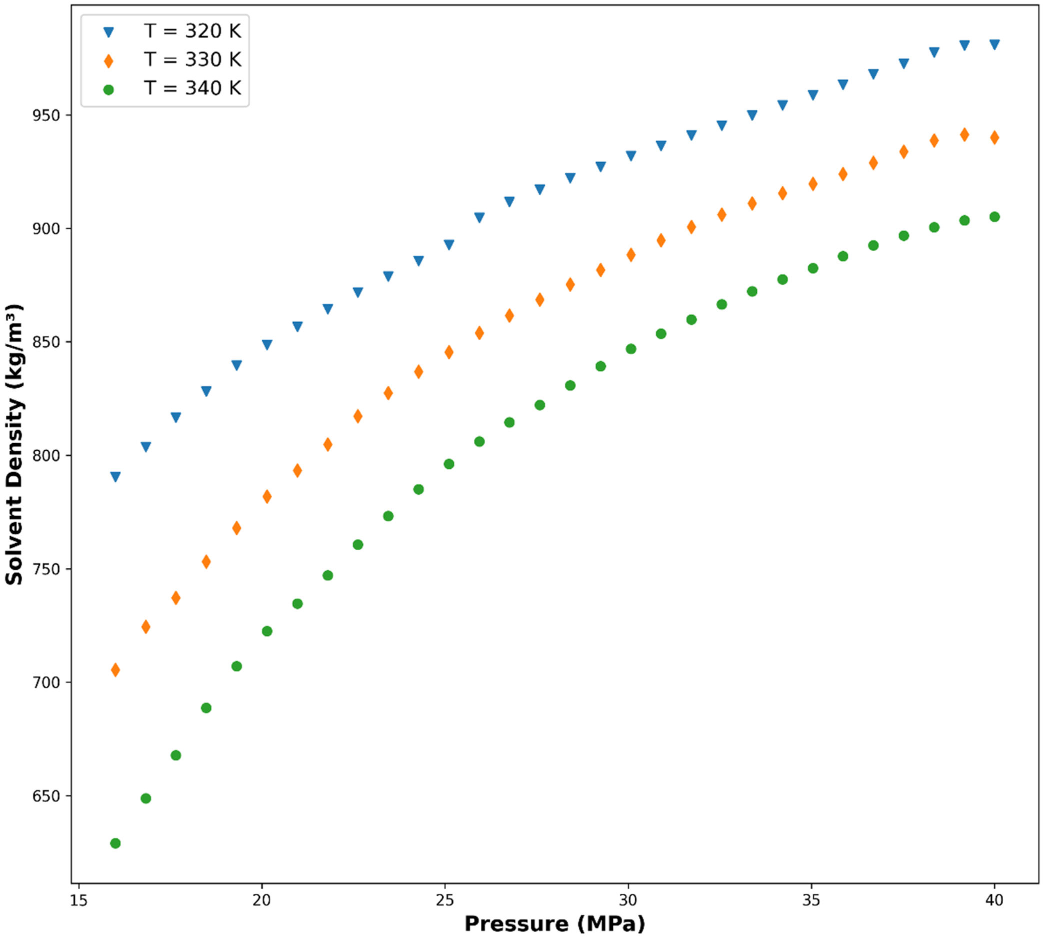

Fig. 13. Effect of pressure on density changes.

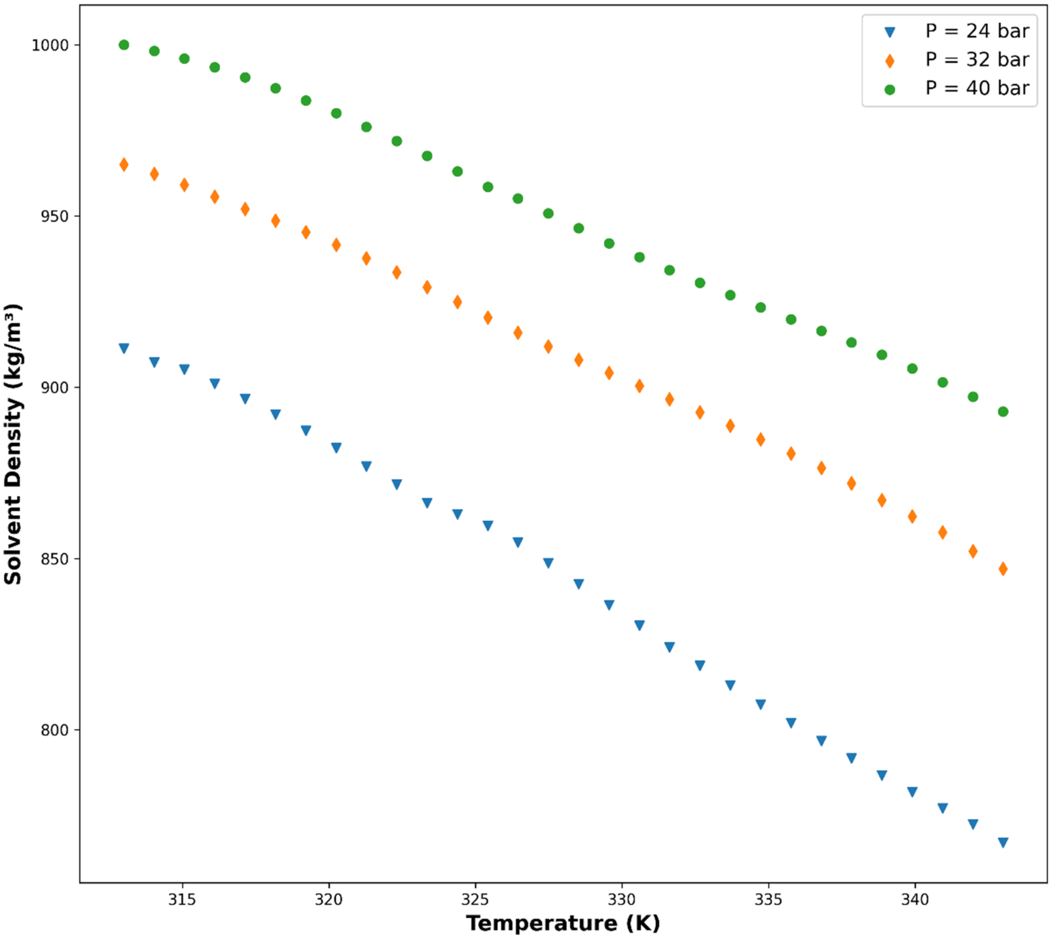

Fig. 14. Effect of temperature on density changes.

Data availability

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

Received: 21 May 2025; Accepted: 4 December 2025
Published online: 10 December 2025

References

1. Abuhassan, Q. et al. A novel simulated media system for in vitro evaluation of bioequivalent intestinal drug solubility. *Eur. J. Pharm. Biopharm.* **199**, 114302 (2024).
2. Cenci, F. et al. Predicting drug solubility in organic solvents mixtures: A machine-learning approach supported by high-throughput experimentation. *Int. J. Pharm.* 124233 (2024).
3. de Waal, T. et al. Characterization of neonatal and infant enterostomy fluids - part II: drug solubility. *Int. J. Pharm.* **642**, 123141 (2023).
4. Fazel-Hoseini, S. M. et al. Modeling of drug solubility with extended Hildebrand solubility approach and jouyban-acree equations in binary and ternary solvent mixtures. *J. Drug Deliv. Sci. Technol.* **95**, 105634 (2024).
5. Lahiq, A. A., Alshehri, A. A. & Alsharif, S. T. Machine learning analysis of drug solubility via green approach to enhance drug solubility for poor soluble medications in continuous manufacturing. *Sci. Rep.* **15** (1), 26007 (2025).
6. 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).
7. AravindKumar, P. et al. New solubility model to correlate solubility of anticancer drugs in supercritical carbon dioxide and evaluation with Kruskal–Wallis test. *Fluid. Phase. Equilibria.* **582**, 114099 (2024).
8. Jouyban, A. A global model for simulation of drug solubility in mono-solvents at different temperatures. *J. Mol. Liq.* **389**, 122850 (2023).
9. Patel, R. P. et al. In vitro lipolysis model to predict food effect of poorly water-soluble drugs itraconazole, rivaroxaban, and ritonavir. *J. Pharm. Sci.* (2024).
10. Abourehab, M. A. S. et al. Enhancing drugs bioavailability using nanomedicine approach: predicting solubility of Tolmetin in supercritical solvent via advanced computational techniques. *J. Mol. Liq.* **365**, 120103 (2022).
11. Ghazwani, M. & Yasmin Begum, M. Machine learning aided drug development: assessing improvement of drug efficiency by correlation of solubility in supercritical solvent for nanomedicine preparation. *J. Mol. Liq.* **387**, 122511 (2023).
12. Hao, C. et al. Computational study and experimental validation on the solubility of drugs in supercritical solvent for assessment of nanomedicine production via green technology for enhanced drug bioavailability. *J. Mol. Liq.* **382**, 121835 (2023).

13. Luo, B. et al. Experimental validation and modeling study on the drug solubility in supercritical solvent: case study on exemestane drug. *J. Mol. Liq.* **377**, 121517 (2023).
14. Yang, F. et al. Artificial intelligence for computation and development of nanodrug solubility in supercritical solvent: analysis of temperature and pressure influence. *J. Mol. Liq.* **414**, 126095 (2024).
15. Liu, C. et al. An improved anticancer Drug-Response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Therapy Nucleic Acids.* **21**, 676–686 (2020).
16. Zhong, S. et al. Machine learning: New ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **55**(19), 12741–12754 (2021).
17. Zhou, Z. H. *Machine Learning* (Springer Nature, 2021).
18. Collins, M., Schapire, R. E. & Singer, Y. Logistic regression, adaboost and Bregman distances. *Mach. Learn.* **48**, 253–285 (2002).
19. Yang, M., Yue, Q. & He, J. Evaluation of mesoporous silica synthesized for green adsorption by modeling via machine learning and mass transfer. *Sci. Rep.* **15** (1), 19477 (2025).
20. Groß, J. *Linear Regression*, vol. 175 (Springer Science & Business Media, 2003).
21. Schölkopf, B. et al. New support vector algorithms. *Neural Comput.* **12** (5), 1207–1245 (2000).
22. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
23. 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).
24. Zhang, X. Employment of a machine learning-based modeling and simulation to perceive the connections between material properties and quality attributes in pharmaceuticals. *Chin. J. Phys.* (2025).
25. Sumayli, A., Mahdi, W. A. & Alamoudi, J. A. Analysis of nanomedicine production via green processing: modeling and simulation of pharmaceutical solubility using artificial intelligence. *Case Stud. Therm. Eng.* **51**, 103587 (2023).
26. Khishe, M. & Mosavi, M. R. Chimp optimization algorithm. *Expert Syst. Appl.* **149**, 113338 (2020).
27. Alorf, A. A survey of recently developed metaheuristics and their comparative analysis. *Eng. Appl. Artif. Intell.* **117**, 105622 (2023).
28. Jia, H. et al. An enhanced chimp optimization algorithm for continuous optimization domains. *Complex. Intell. Syst.* 1–18 (2021).
29. Ying, C. et al. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica.* **39** (6), 745–758 (2013).
30. James, G. et al. Linear regression. In *An Introduction To Statistical Learning: with Applications in Python*, 69–134 (Springer, 2023).
31. Zhang, F. & O'Donnell, L. J. Support vector regression. In *Machine Learning*, 123–140 (Elsevier, 2020).
32. Wang, J. *An intuitive tutorial to Gaussian processes regression.* arXiv preprint arXiv:2009.10862 (2020).
33. Pourhanasa, R. & Monadipour, A. Concrete crack detection via graph representation learning and texture analysis. *Innov. Infrastr. Solut.* **10** (8), 1–12 (2025).

Acknowledgements

The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2025-24).

Author contributions

A.A.L.: Developed the model, wrote manuscript, performed modeling, provided supervision. A.A.A.: Performed validation, wrote the manuscript, developed methodology. S.T.A.: Wrote the manuscript, conducted visualization, performed simulation, performed validation. All authors reviewed the manuscript.

Funding

This research was funded by Taif University, Saudi Arabia, Project No. (TU-DSPP-2025-24).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

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

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.