molecules

MDPI

Article

Anti-Cancer Drug Solubility Development within a Green Solvent: Design of Novel and Robust Mathematical Models Based on Artificial Intelligence

Bader Huwaimel [1,*] and Ahmed Alobaida [2]

[1] Department of Pharmaceutical Chemistry, College of Pharmacy, University of Hail, Hail 81442, Saudi Arabia
[2] Depaertmen of Pharmaceutics, College of Pharmacy, University of Hail, Hail 81442, Saudi Arabia
* Correspondence: b.huwaimel@uoh.edu.sa

Abstract: Nowadays, supercritical CO_2(SC-CO_2) is known as a promising alternative for challengeable organic solvents in the pharmaceutical industry. The mathematical prediction and validation of drug solubility through SC-CO_2 system using novel artificial intelligence (AI) approach has been considered as an interesting method. This work aims to evaluate the solubility of tamoxifen as a chemotherapeutic drug inside the SC-CO_2 via the machine learning (ML) technique. This research employs and boosts three distinct models utilizing Adaboost methods. These models include K-nearest Neighbor (KNN), Theil-Sen Regression (TSR), and Gaussian Process (GPR). Two inputs, pressure and temperature, are considered to analyze the available data. Furthermore, the output is Y, which is solubility. As a result, ADA-KNN, ADA-GPR, and ADA-TSR show an R^2 of 0.996, 0.967, 0.883, respectively, based on the analysis results. Additionally, with MAE metric, they had error rates of 1.98×10^{-6}, 1.33×10^{-6}, and 2.33×10^{-6}, respectively. A model called ADA-KNN was selected as the best model and employed to obtain the optimum values, which can be represented as a vector: (X1 = 329, X2 = 318.0, Y = 6.004×10^{-5}) according to the mentioned metrics and other visual analysis.

Keywords: pharmaceutical industry; supercritical CO_2; drug solubility; predictive models

Citation: Huwaimel, B.; Alobaida, A. Anti-Cancer Drug Solubility Development within a Green Solvent: Design of Novel and Robust Mathematical Models Based on Artificial Intelligence. *Molecules* **2022**, *27*, 5140. https://doi.org/10.3390/molecules27165140

Academic Editor: Young Hae Choi

Received: 18 June 2022
Accepted: 9 August 2022
Published: 12 August 2022

1. Introduction

The discovery of novel drug molecules followed by their introduction into clinical trials is considered as the main goal of the drug development industry for increasing the efficiency and reducing the side effects of drugs [1–4]. Solubility is one of the main parameters that influence drug efficiency [5,6]. Low solubility is considered as the most important challenge towards the formulation of novel chemical entities [7]. Various techniques can be used to improve drug solubility, such as physical modification (i.e., nanosuspension), chemical modification (i.e., complexation and salt formation), and miscellaneous procedures (i.e., supercritical fluids (SCFs) process and solubilizers) [8–11].

SCFs (especially supercritical CO_2 (SC-CO_2)) have been recently identified as a promising alternative for challengeable organic solvents. The emergence of remarkably positive points such as cost-effectiveness, inert nature, environmentally friendly, excellent chemical affinity in almost all organic solvents, safety of application and non-toxic characteristic has improved the tendency of researchers to apply them in pharmacology [12–14]. Additionally, the modulation of two momentous properties of CO_2 including density and solvent power is feasible by the alteration of operational pressure/temperature and true control of the process kinetics [15–18].

The development of predictive models to estimate the solubility of various types of drugs in real conditions has been an interesting topic. Artificial intelligence (AI) approach is known as a robust and efficient approach to mathematically predict the results in various scientific scopes, such as nanotechnology, separation, extraction, chemical reactors, and transport phenomena [19–23].

Machine learning (ML) is a set of techniques and tools that uses data to create a mathematical model to make predictions or perform analysis, and it is critical in artificial intelligence [24,25]. ML approaches are progressively replacing computational methods in scientific domains. ML models may now investigate any problem with several input features and at least one target. These models extract inputs–outputs relationships using various strategies [26–28].

Boosting is a subtype of ensemble techniques that integrate the outcomes of several weak estimators to build a robust estimator. Boosting makes the usage of weak estimators applying a sequential logic, which implies the results of each weak estimate the influence of the following estimate. AdaBoost [29], in particular, is a representative boosting learning method that generates weak estimators gradually utilizing reweighted training data.

In recent years, GPR has gained popularity as a data-driven modeling tool. GPR's popularity stems in part from its theoretical connection to Bayesian nonparametric statistics, infinite neural networks, kernel approaches in machine learning, and spatial statistics [30,31].

If the target data are numeric and continuous, neighbors-based regression such as KNN can be used. A query point's label is determined by averaging the labels of its nearest neighbors [32].

Theil-Sen Regression is another weak estimator is used here. Compared to Ordinary Least Squares (OLS), Theil-Sen Regressor has a comparable asymptotic efficiency and is an unbiased estimate. Since it makes no assumptions about the underlying distribution of the data, Theil-Sen is non-parametric in comparison to OLS. Theil-Sen can withstand outliers more effectively [33,34].

The main novelty of this paper is to predict the optimized value of tamoxifen solubility in an SC-CO_2 system via the ML approach. To achieve this, three ML-based predictive models including K-nearest Neighbor (KNN), Theil-Sen Regression (TSR), and Gaussian Process (GPR) were developed. The comparison of the models showed the fact that ADA-KNN is the most accurate and general model due to more proximity of points with actual test and train data lines and greater R^2 value.

2. Dataset

In this research, a small dataset containing two inputs composed of X1 = P (bar) and X2 = T (K) and the only possible output is Y = solubility was applied. There are only 32 data points that were taken from the literature, and they performed the analysis for the pressure of 120–400 bar and temperature of 308–338 K [35]. The entire dataset is displayed in Table 1.

Table 1. Dataset.

No.	X1 = P (bar)	X2 = T (K)	Y (Solubility/Mole Fraction)
1	120	308	4×10^{-6}
2	160	308	4.94×10^{-6}
3	200	308	5.49×10^{-6}
4	240	308	5.96×10^{-6}
5	280	308	3.99×10^{-6}
6	320	308	3.88×10^{-6}
7	360	308	8.38×10^{-6}
8	400	308	1.24×10^{-5}
9	120	318	2.15×10^{-6}
10	160	318	5.79×10^{-6}
11	200	318	8.95×10^{-6}
12	240	318	7.27×10^{-6}

Table 1. *Cont.*

No.	X1 = P (bar)	X2 = T (K)	Y (Solubility/Mole Fraction)
13	280	318	3.40×10^{-6}
14	320	318	7.03×10^{-5}
15	360	318	4.01×10^{-6}
16	400	318	1.39×10^{-5}
17	120	328	1.79×10^{-6}
18	160	328	5.13×10^{-6}
19	200	328	1.05×10^{-6}
20	240	328	5.48×10^{-5}
21	280	328	2.31×10^{-5}
22	320	328	2.04×10^{-5}
23	360	328	2.50×10^{-5}
24	400	328	4.41×10^{-5}
25	120	338	1.52×10^{-5}
26	160	338	3.84×10^{-6}
27	200	338	1.05×10^{-5}
28	240	338	2.08×10^{-5}
29	280	338	3.13×10^{-5}
30	320	338	1.95×10^{-5}
31	360	338	5.47×10^{-5}
32	400	338	6.0×10^{-5}

3. Methodology

3.1. Base Models

The first base model is a kernel-based and non-parametric method, Gaussian process regression (GPR). GPR focused on statistical learning theory and Bayesian models. When used in conjunction with the mean function, a kernel can be used to explain the covariance function of a Gaussian random variable. The GPR's capacity to generalize well, particularly when working with minor data sets, is one of its most significant advantages [36–38]. When constructing a GPR model, the following equation is assumed to be true for an output Y:

$$Y = f(X) + \xi \tag{1}$$

f(X) illustrates the underlying function, X as input of the training data, X_* as test subset, and $\xi \sim N(0, \sigma^2)$ as the error. The error variance σ^2 is calculated based on the input vector. The previous joint distribution of the actual target Y and the expected target y are [39,40]:

$$Y \sim N\left(0, K(X,X) + \sigma^2 I\right) \tag{2}$$

$$\begin{bmatrix} Y \\ y \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) = N\left(\begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{3}$$

$K = (k_{ij})$ as the covariance kernel matrix of the train subset in which the elements measure the relation between X_i and X_j through k. K_* stands for the covariance matrix between the test and train subsets, and K_{**} indicates the covariance matrix of the test subset [36,41]. The

posterior distribution (in Bayesian analysis, reflects information about uncertain quantities) of y is shown in Equations (4)–(6):

$$y \Big| Y \sim N\left(\overline{y}, \sigma_y^2\right) \tag{4}$$

$$\overline{y} = K_* K^{-1} Y \tag{5}$$

$$\sigma_y^2 = K_{**} - K_* K^{-1} K_*^T \tag{6}$$

The other base models are K-nearest neighbor regression (KNN). The KNN regressor learns by comparison of the identified test examples to the training set [42]. $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ represent the training data with a parameter of distance d. $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im})$ represent the i-th sample indicates with m input features and its target output y_i. Additionally, N represent the count of examples. It must calculate the d_i between a test instance x and any sample $x_i \in T$ and sort the d_i distance by its value for a test sample x. If d_i is in the i-th place, the instance of x matches di, which is called the i-th nearest neighbor, or $NN_i(x)$, and its target is called $y_i(x)$. Lastly, the estimation $\hat{y}$ of input instance x denotes the average of the prediction of k-nearest neighbors to x ($\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i(x)$). KNN regression algorithm can be summarized in the following steps [43]:

- Inputs: training samples $\{x_i, y_i\}$, x_i: input features, y_i: real-valued output, testing point x to predict
- Algorithm:
- Calculate distance $D(x, x_i)$ to every training example x_i
- Select k closet examples $x_{i1} \ldots x_{ik}$ and their outputs $y_{i1} \ldots y_{ik}$
- Output:

$$\hat{y} = f(x) = \frac{1}{K} \sum_{j=1}^{k} y_{i_j} \tag{7}$$

The third base model is Theil-Sen Regression. The model is estimated in Theil-Sen regression by computing the slopes and intercepts of a subset of all feasible solutions of p subsample points. When an intercept is fitting, p must be bigger than or equal to number of features + 1. The spatial median of these slopes and intercepts is then used to define the final slope and intercept.

The trend slopes were estimated using the Theil-Sen (TSR) estimator [44], which was chosen since it is better than raw linear regression approaches in evaluating trend slopes in the existence of outliers in data [45].

The initial phase in calculating the TSR predictor is to determine the Q_i value given N pairs of data [44]:

$$Q_i = \frac{x_j - x_k}{j - k} \; i \in \{1, 2, \ldots, N\} \tag{8}$$

x_j, x_k are the data point vectors.

If only one datum is existed, then $N = \frac{n(n-1)}{2}$. Additionally, n is the count of vectors. If there are many observed data in several vectors, then $N < \frac{n(n-1)}{2}$, n is the count of observed vectors.

Then, the TSR predictor is calculated as the median Q_{med} of the N values of Q_i, sorted in (minimum, maximum) interval [44]:

$$Q_{med} = \begin{cases} Q_{\frac{(N+1)}{2}} \; when \; N \; is \; odd \\ \frac{Q_{\frac{N}{2}} + Q_{\frac{(N+1)}{2}}}{2} \; when \; N \; is \; even \end{cases} \tag{9}$$

The sign of Q_{med} shows the trend behavior, and its value shows the magnitude of the trend.

3.2. AdaBoost

Adaboost [46] is the most well-known boosting model, and it was initially employed to address the classification issue. Freund [29] then presented the Adaboost.R to handle real-valued regression problems. Additionally, drucker [47] solved the regression problem using the updated Adaboost.R2 model, with amazing results.

The data sample weights are set to zero. The initial iteration trains a weak learner, and the instance weights are adjusted based on the training outcomes. The adjusted weights are used to train the next weak learner. Each iteration, the weights of the instances estimated incorrectly (with a high error) in the previous iteration are increased, while the weights of the instances estimated correctly (near expected value) in the previous iteration are decreased. The influence of hard-to-predict instances becomes increasingly substantial as the number of iterations grows; after each iteration, the weak learner concentrates more on samples that were previously estimated poorly. The final prediction outcome is established by a weak learner's weighted vote. Any machine learning regression technique may be used to choose the weak learner in AdaBoost regression [48–50]. In this study, we used three models of previous section as weak learners distinctly.

4. Results

We employed grid search to find the optimal hyper-parameters of these models and obtained the final configuration of each model. MAE and R^2 are two metrics that were used to evaluate the performance of the model that were calculated using Equations (10) and (11) [51,52].

$$\mathrm{MAE} = \frac{1}{n}\sum\left|x_i^{t+1} - x_i^{t+1}\right| \tag{10}$$

$$\mathrm{R2} = 1 - \frac{\sum\left(\hat{x}_i^{t+1} - x_i^{t+1}\right)^2}{\sum\left(x_i^{t+1} - x_i^{t+1}\right)^2} \tag{11}$$

In these equations, x_i^{t+1} is the estimated value, x_i^{t+1} is the observed value, and n is the quantity of examples.

The accuracy of the final models is presented in Table 2 Additionally, the comparison of expected and estimated values of tamoxifen solubility in SC-CO_2 system via ADA-KNN, ADA-GPR, and ADA-TSR models is shown in Figures 1–3. In these diagrams, the green line is the actual data line, and the point is predicted values blue for train sunset and red for test subset. Comparing these three charts proves that the ADA-KNN is the most general and appropriate model since the points are near actual test and train data lines.

Table 2. Output.

Models	MAE	R^2
ADA-KNN	1.98×10^{-6}	0.996
ADA-GPR	1.33×10^{-6}	0.967
ADA-TSR	2.33×10^{-6}	0.883

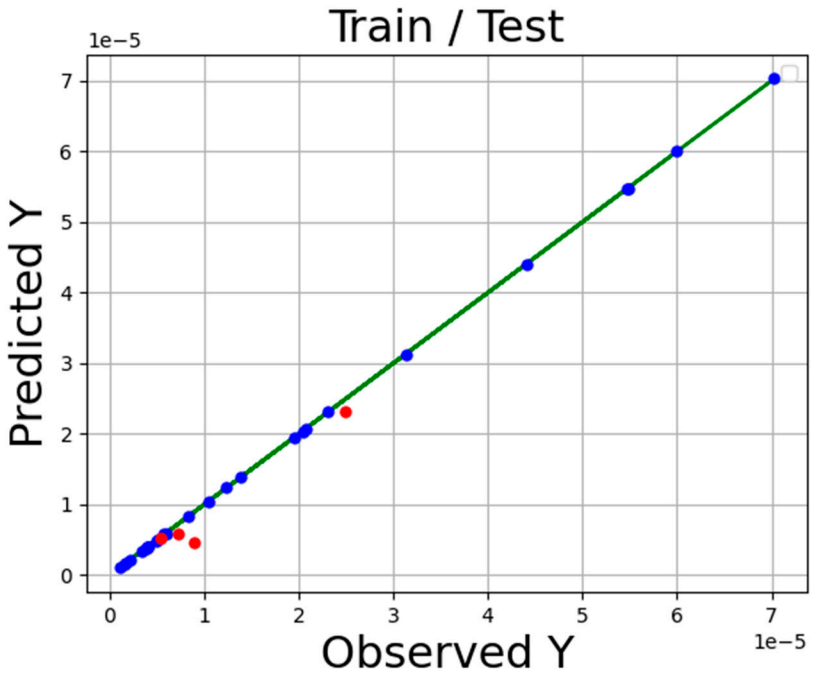

Figure 1. Fitting chart for ADA-KNN.

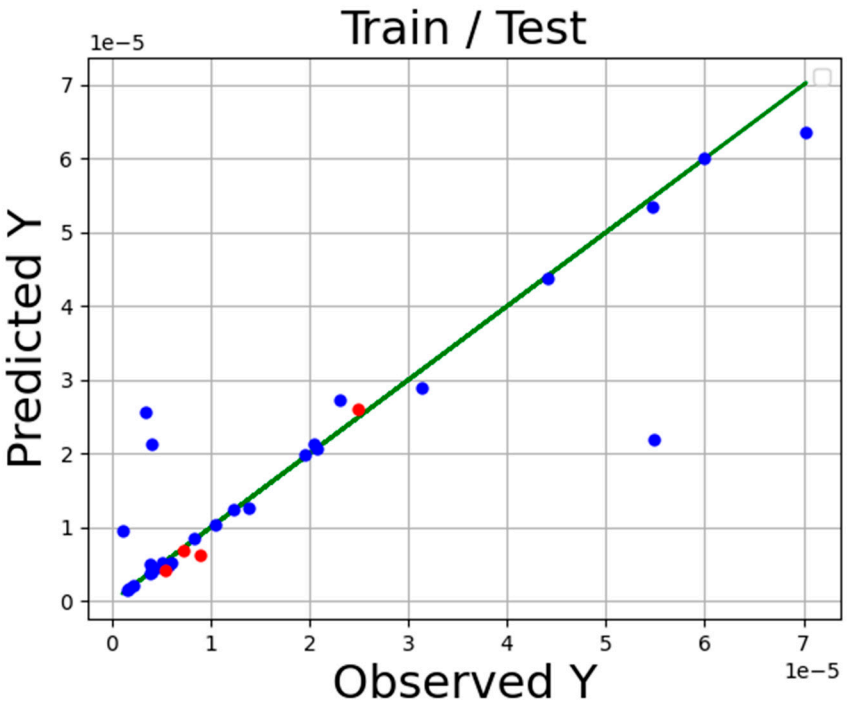

Figure 2. Fitting chart for ADA-GPR.

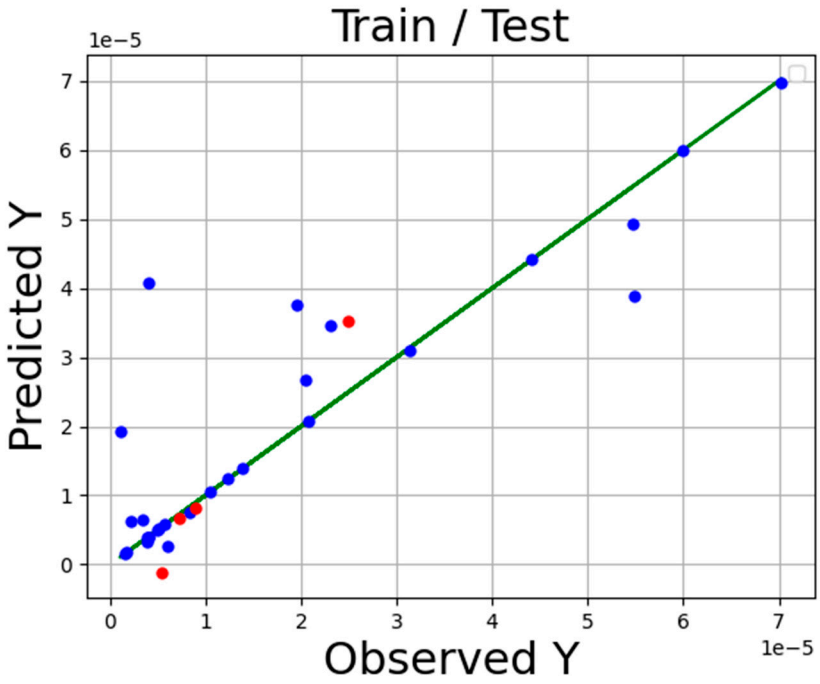

Figure 3. Fitting chart for ADA-TSR.

Figure 4 illustrates the three-dimensional projection to demonstrate the final results of the ADA-KNN mathematical model to measure the impacts of input parameters (pressure and temperature) on drug solubility at the same time. Furthermore, two-dimensional depictions to individually evaluate the effects of pressure and temperature on the values of tamoxifen solubility in SC-CO_2 system are shown in Figures 5 and 6. It can be seen from the figures that pressure has positive effect on the solubility value of drugs in the SC-CO_2 fluid due to increasing the density of SCFs owing to modify the molecular compaction. If the value of density increases, the solvating capability of solvent increases significantly and, the solubility of drug in SC-CO_2 increases. The effect of temperature on drug solubility is paradoxical. In one side, increment of temperature improves the pressure sublimation of solvent, which is a positive phenomenon in increasing the solubility of the drug inside SCFs. On the other side, the increase in temperature reduces the density of solvent, which considerably deteriorates the solvating power and consequently solubility amount of drug. Considering the abovementioned explanations, the net impacts of the sublimation pressure and density can determine the favorable/unfavorable role of temperature on the solubility. The evaluation of figures illustrates the emergence of a cross-over pressure in the isotherms. At the pressures over than cross-over pressure, an increase in temperature improves the drug solubility because of the greater effect of sublimation pressure compared to density. For the pressures lower than the cross-over pressure, increasing the temperature, decrement in the solvent density overcomes the effect of pressure sublimation and as a result, and decreases the tamoxifen solubility in SC-CO_2 fluid [35]. Based on the presented results of Table 3, the pressure and temperature at 329 bar and 318 K, respectively, were considered as the optimum pressure and temperature for reaching the maximum amount of tamoxifen solubility.

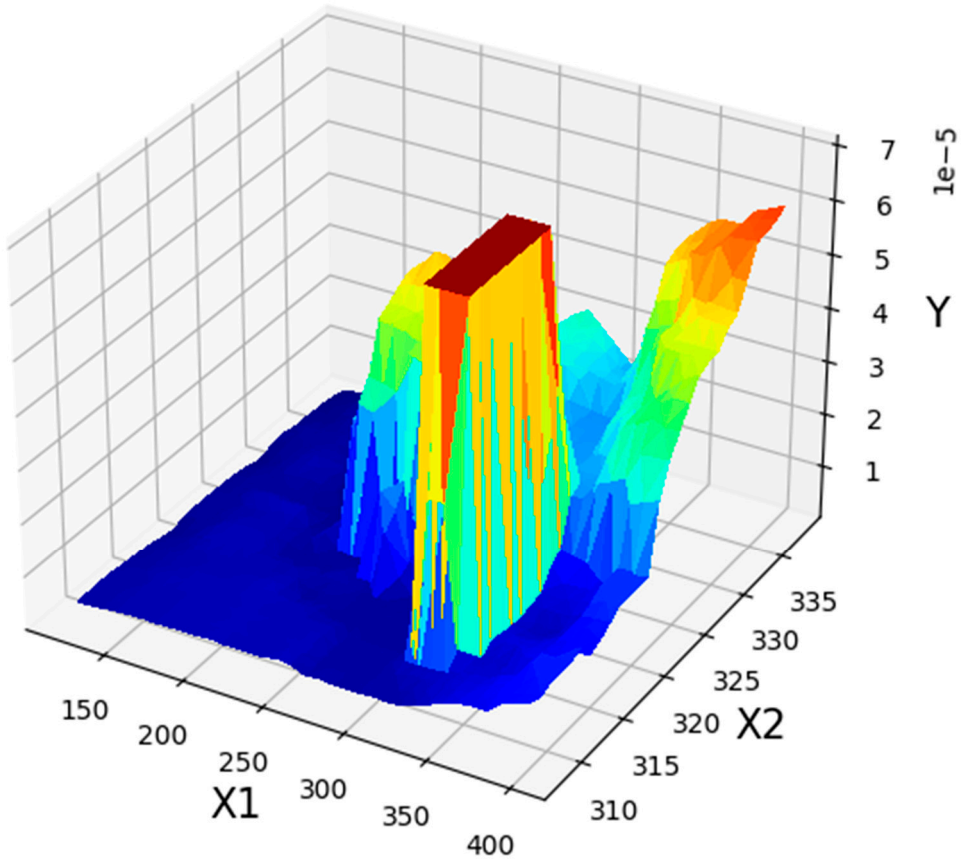

Figure 4. Three-dimensional illustration of pressure (X1), temperature (X2), and solubility (Y).

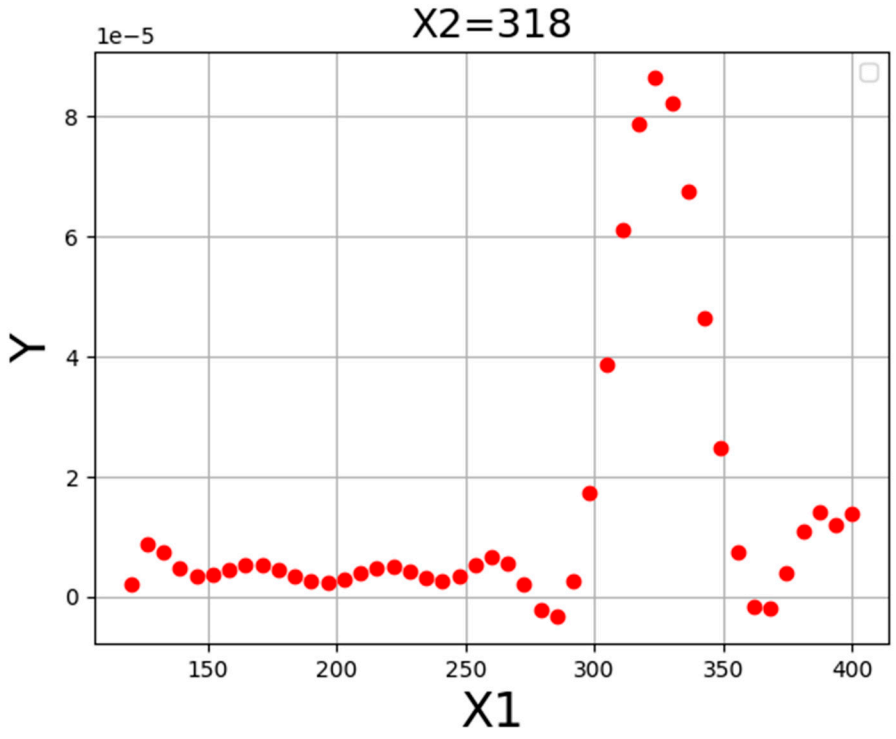

Figure 5. Tendency of X1.

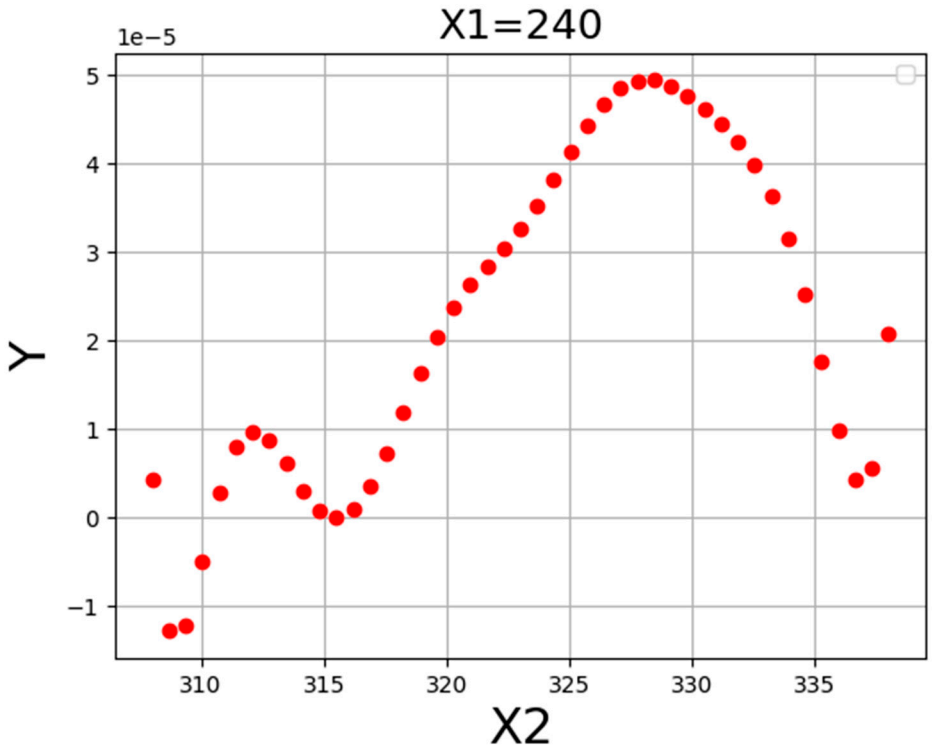

Figure 6. Tendency of X2.

Table 3. Modified parameters applying maximum response.

X1 = P (bar)	X2 = T (K)	Y (Solubility)
329	318.0	7.03×10^{-5}

5. Conclusions

In this research work, three new models were compared through machine learning to estimate and validate the solubility of tamoxifen in supercritical CO_2. The Adaboost method was applied to improve these three different models, including KNN, GPR and TSR, and the results are promising. According to the analysis, the R^2 of the ADA-KNN, ADA-GPR, and ADA-TSR models were 0.996, 0.967, and 0.883, respectively. The MAE error rates for these three models were 1.98×10^{-6}, 1.33×10^{-6}, and 2.33×10^{-6}, respectively. An ADA-KNN model was selected as the best model, and it was applied to optimize the values using these metrics (X1 = 329, X2 = 318.0, Y = 6.004×10^{-5}) and some visual analysis.

Author Contributions: B.H.: Writing, editing, data analysis, methodology, conceptualization, validation, resources, supervision, data cura-tion, A.A.: review and editing, validation, resources, visualization, software. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available within the published paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Faqi, A.S. *A Comprehensive Guide to Toxicology in Nonclinical Drug Development*; Academic Press: London, UK, 2016.
2. Brun, R.; Don, R.; Jacobs, R.T.; Wang, M.Z.; Barrett, M.P. Development of novel drugs for human African trypanosomiasis. *Future Microbiol.* **2011**, *6*, 677–691. [CrossRef]
3. Martell, R.E.; Brooks, D.G.; Wang, Y.; Wilcoxen, K. Discovery of novel drugs for promising targets. *Clin. Ther.* **2013**, *35*, 1271–1281. [CrossRef]
4. Mirhaji, E.; Afshar, M.; Rezvani, S.; Yoosefian, M. Boron nitride nanotubes as a nanotransporter for anti-cancer docetaxel drug in water/ethanol solution. *J. Mol. Liq.* **2018**, *271*, 151–156. [CrossRef]
5. Savjani, K.T.; Gajjar, A.K.; Savjani, J.K. Drug solubility: Importance and enhancement techniques. *Int. Sch. Res. Not.* **2012**, *2012*, 195727. [CrossRef] [PubMed]
6. Gorain, B.; Pandey, M.; Choudhury, H.; Jain, G.K.; Kesharwani, P. Dendrimer for solubility enhancement. In *Dendrimer-Based Nanotherapeutics*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 273–283.
7. Williams, H.D.; Trevaskis, N.L.; Charman, S.A.; Shanker, R.M.; Charman, W.N.; Pouton, C.W.; Porter, C.J.H. Strategies to address low drug solubility in discovery and development. *Pharmacol. Rev.* **2013**, *65*, 315–499. [CrossRef]
8. Vimalson, D.C. Techniques to enhance solubility of hydrophobic drugs: An overview. *Asian J. Pharm.* **2016**, 10. [CrossRef]
9. Das, B.; Baidya, A.T.; Mathew, A.T.; Yadav, A.K.; Kumar, R. Structural modification aimed for improving solubility of lead compounds in early phase drug discovery. *Bioorganic Med. Chem.* **2022**, *56*, 116614. [CrossRef] [PubMed]
10. Bagade, O.; Kad, D.R.; Bhargude, D.N.; Bhosale, D.R.; Kahane, S.K. Consequences and impose of solubility enhancement of poorly water soluble drugs. *Res. J. Pharm. Technol.* **2014**, *7*, 598.
11. Cao, M.; Yoosefian, D.W.M.; Sabaei, S.; Jahani, M. Comprehensive study of the encapsulation of Lomustine anticancer drug into single walled carbon nanotubes (SWCNTs): Solvent effects, molecular conformations, electronic properties and intramolecular hydrogen bond strength. *J. Mol. Liq.* **2020**, *320*, 114285. [CrossRef]
12. Girotra, P.; Singh, S.K.; Nagpal, K. Supercritical fluid technology: A promising approach in pharmaceutical research. *Pharm. Dev. Technol.* **2013**, *18*, 22–38. [CrossRef]
13. Macnaughton, S.J.; Kikic, I.; Foster, N.R.; Alessi, P.; Cortesi, A.; Colombo, I. Solubility of anti-inflammatory drugs in supercritical carbon dioxide. *J. Chem. Eng. Data* **1996**, *41*, 1083–1086. [CrossRef]
14. Zhou, M.; Ni, R.; Zhao, Y.; Huang, J.; Deng, X. Research progress on supercritical CO_2 thickeners. *Soft Matter* **2021**, *1*, 5107–5115. [CrossRef] [PubMed]
15. Baldino, L.; Cardea, S.; Reverchon, E. Biodegradable membranes loaded with curcumin to be used as engineered independent devices in active packaging. *J. Taiwan Inst. Chem. Eng.* **2017**, *71*, 518–526. [CrossRef]
16. Su, W.; Zhang, H.; Xing, Y.; Li, X.; Wang, J.; Cai, C. A bibliometric analysis and review of supercritical fluids for the synthesis of nanomaterials. *Nanomaterials* **2021**, *11*, 336. [CrossRef]
17. Baldino, L.; della Porta, G.; Reverchon, E. Supercritical CO_2 processing strategies for pyrethrins selective extraction. *J. CO2 Util.* **2017**, *20*, 14–19. [CrossRef]
18. Yoosefian, M.; Sabaei, S.; Etminan, N. Encapsulation efficiency of single-walled carbon nanotube for Ifosfamide anti-cancer drug. *Comput. Biol. Med.* **2019**, *114*, 103433. [CrossRef] [PubMed]
19. Zhu, H.; Zhu, L.; Sun, Z.; Khan, A. Machine learning based simulation of an anti-cancer drug (busulfan) solubility in supercritical carbon dioxide: ANFIS model and experimental validation. *J. Mol. Liq.* **2021**, *338*, 116731. [CrossRef]
20. Öztürk, A.A.; Gündüz, A.B.; Ozisik, O. Supervised machine learning algorithms for evaluation of solid lipid nanoparticles and particle size. *Comb. Chem. High Throughput Screen.* **2018**, *21*, 693–699. [CrossRef] [PubMed]
21. Staszak, M. Artificial intelligence in the modeling of chemical reactions kinetics. *Phys. Sci. Rev.* **2020**. [CrossRef]
22. Wang, X.; Luo, L.; Xiang, J.; Zheng, S.; Shittu, S.; Wang, Z.; Zhao, X. A comprehensive review on the application of nanofluid in heat pipe based on the machine learning: Theory, application and prediction. *Renew. Sustain. Energy Rev.* **2021**, *150*, 111434. [CrossRef]
23. Lazzús, J.A.; Cuturrufo, F.; Pulgar-Villarroel, G.; Salfate, I.; Vega, P. Estimating the temperature-dependent surface tension of ionic liquids using a neural network-based group contribution method. *Ind. Eng. Chem. Res.* **2017**, *56*, 6869–6886. [CrossRef]
24. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
25. Mitchell, T.M. *The Discipline of Machine Learning*; Carnegie Mellon University: Pittsburgh, PA, USA, 2006; Volume 9.
26. El Naqa, I.; Murphy, M.J. What is machine learning? In *Machine Learning in Radiation Oncology*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 3–11.
27. Goodfellow, I.; Bengio, Y.; Courville, A. Machine learning basics. *Deep. Learn.* **2016**, *1*, 98–164.
28. Shehadeh, A.; Alshboul, O.; Al Mamlook, R.E.; Hamedat, O. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Autom. Constr.* **2021**, *129*, 103827. [CrossRef]
29. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
30. Rasmussen, C.E. Gaussian processes in machine learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003.
31. Shi, J.Q.; Choi, T. *Gaussian Process Regression Analysis for Functional Data*; CRC Press: Boca Raton, FL, USA, 2011.

32. Masegosa, R.A.; Armañanzas, R.; Abad-Grau, M.M.; Potenciano, V.; Moral, S.; Larrañaga, P.; Bielza, C.; Matesanz, F. Discretization of Expression Quantitative Trait Loci in Association Analysis Between Genotypes and Expression Data. *Curr. Bioinform.* **2015**, *10*, 144–164. [CrossRef]
33. Wilcox, R. A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biom. J.* **1998**, *40*, 261–268. [CrossRef]
34. Ohlson, J.A.; Kim, S. Linear valuation without OLS: The Theil-Sen estimation approach. *Rev. Account. Stud.* **2015**, *20*, 395–435. [CrossRef]
35. Pishnamazi, M.; Zabihi, S.; Jamshidian, S.; Borousan, F.; Hezave, A.Z.; Shirazian, S. Thermodynamic modelling and experimental validation of pharmaceutical solubility in supercritical solvent. *J. Mol. Liq.* **2020**, *319*, 114120. [CrossRef]
36. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Regression*; 1996.
37. Rasmussen, C.E. *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*; University of Toronto: Toronto, ON, Canada, 1997.
38. Taherdangkoo, R.; Yang, H.; Akbariforouz, M.; Sun, Y.; Liu, Q.; Butscher, C. Gaussian process regression to determine water content of methane: Application to methane transport modeling. *J. Contam. Hydrol.* **2021**, *243*, 103910. [CrossRef]
39. Alghamdi, A.S.; Polat, K.; Alghoson, A.; Alshdadi, A.A.; Abd El-Latif, A.A. Gaussian process regression (GPR) based non-invasive continuous blood pressure prediction method from cuff oscillometric signals. *Appl. Acoust.* **2020**, *164*, 107256. [CrossRef]
40. Cheng, M.; Prayogo, D. *Optimizing Biodiesel Production from Rice Bran Using Artificial Intelligence Approaches*; Department of Construction Engineering, National Taiwan University of Science and Technology: Taipei, Taiwan, 2016.
41. Williams, C.K.; Barber, D. Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1342–1351. [CrossRef]
42. Cover, T. Estimation by the nearest neighbor rule. *IEEE Trans. Inf. Theory* **1968**, *14*, 50–55. [CrossRef]
43. Song, Y.; Liang, J.; Lu, J.; Zhao, X. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing* **2017**, *251*, 26–34. [CrossRef]
44. Sen, P.K. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* **1968**, *63*, 1379–1389. [CrossRef]
45. Caloiero, T.; Aristodemo, F.; Ferraro, D.A. Annual and seasonal trend detection of significant wave height, energy period and wave power in the Mediterranean Sea. *Ocean. Eng.* **2022**, *243*, 110322. [CrossRef]
46. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; Citeseer: Princeton, NJ, USA, 1996.
47. Drucker, H. Improving regressors using boosting techniques. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 1997; Citeseer: Princeton, NJ, USA, 1997.
48. Dargahi-Zarandi, A.; Hemmati-Sarapardeh, A.; Shateri, M.; Menad, N.A.; Ahmadi, M. Modeling minimum miscibility pressure of pure/impure CO2-crude oil systems using adaptive boosting support vector regression: Application to gas injection processes. *J. Pet. Sci. Eng.* **2020**, *184*, 106499. [CrossRef]
49. Wu, Q.; Burges, C.J.C.; Svore, K.M.; Gao, J. Adapting boosting for information retrieval measures. *Inf. Retr.* **2010**, *13*, 254–270. [CrossRef]
50. Ying, C.; Miao, Q.; Liu, J.; Gao, L. Advance and prospects of AdaBoost algorithm. *Acta Autom. Sin.* **2013**, *39*, 745–758.
51. Botchkarev, A. Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio. 2018. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3177507 (accessed on 9 August 2022).
52. Kumar, S.; Mishra, S.; Singh, S.K. A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere. *Heliyon* **2020**, *6*, e05618. [CrossRef]