scientific reports

Check for updates

OPEN

Optimization of tamoxifen solubility in carbon dioxide supercritical fluid and investigating other molecular targets using advanced artificial intelligence models

Saad M. Alshahrani[1]✉, Abdullah S. Alshetaili[1], Munerah M. Alfadhel[1], Amany Belal[2,3]✉, Mohammad A. S. Abourehab[3,4,5], Ahmed Al Saqr[1], Bjad K. Almutairy[1], Kumar Venkatesan[6], Amal M. Alsubaiyel[7]✉ & Mahboubeh Pishnamazi[8,9]✉

Particle size, shape and morphology can be considered as the most significant functional parameters, their effects on increasing the performance of oral solid dosage formulation are indisputable. Supercritical Carbon dioxide fluid ($SCCO_2$) technology is an effective approach to control the above-mentioned parameters in oral solid dosage formulation. In this study, drug solubility measuring is investigated based on artificial intelligence model using carbon dioxide as a common supercritical solvent, at different pressure and temperature, 120–400 bar, 308–338 K. The results indicate that pressure has a strong effect on drug solubility. In this investigation, Decision Tree (DT), Adaptive Boosted Decision Trees (ADA-DT), and Nu-SVR regression models are used for the first time as a novel model on the available data, which have two inputs, including pressure, X1 = P(bar) and temperature, X2 = T(K). Also, output is Y = solubility. With an R-squared score, DT, ADA-DT, and Nu-SVR showed results of 0.836, 0.921, and 0.813. Also, in terms of MAE, they showed error rates of 4.30E−06, 1.95E−06, and 3.45E−06. Another metric is RMSE, in which DT, ADA-DT, and Nu-SVR showed error rates of 4.96E−06, 2.34E−06, and 5.26E−06, respectively. Due to the analysis outputs, ADA-DT selected as the best and novel model and the find optimal outputs can be shown via vector: (x1 = 309, x2 = 317.39, Y1 = 7.03e−05).

In current decades, numerous endeavors have been made to develop new therapeutic medicines and optimize the application of existing drugs[1–5]. One of the most important restrictions towards the development of therapeutic drugs is low drug bioavailability, which is mainly owing to insufficient drug solubilities and low dissolution rate[6]. Therefore, finding promising techniques to enhance and optimize the solubility of drugs is an important method. True recognition of the drug solubility is known as a major necessity for developing the supercritical technology in pharmaceutical processing.

[1]Department of Pharmaceutics, College of Pharmacy, Prince Sattam Bin Abdulaziz University, P.O. Box 173, Al-Kharj 11942, Saudi Arabia. [2]Department of Pharmaceutical Chemistry, College of Pharmacy, Taif University, Taif 21944, Saudi Arabia. [3]Medicinal Chemistry Department, Faculty of Pharmacy, Beni-Suef University, Beni-Suef 62514, Egypt. [4]Department of Pharmaceutics, College of Pharmacy, Umm Al-Qura University, Mecca 21955, Saudi Arabia. [5]Department of Pharmaceutics and Industrial Pharmacy, Faculty of Pharmacy, Minia University, Minia 61519, Egypt. [6]Department of Pharmaceutical Chemistry, College of Pharmacy, King Khalid University, Abha 62529, Saudi Arabia. [7]Department of Pharmaceutics, College of Pharmacy, Qassim University, Buraidah 52571, Saudi Arabia. [8]Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam. [9]The Faculty of Pharmacy, Duy Tan University, Da Nang 550000, Viet Nam. ✉email: Sm.Alshahrani@psau.edu.sa; a.belal@tu.edu.sa; asbiel@qu.edu.sa; mahboubehpishnamazi@duytan.edu.vn

Recently, the use of $SCCO_2$ fluid has been recently of paramount interest in pharmaceutical industry for dissolution of various types of drugs and subsequent nanonization[7–9]. The presence of different advantages such as ease of operation, eco-friendliness, and the non-existence of organic solvents in the production process has increased the interest of scientists to use $SCCO_2$ fluid for enhancing the solubility of drugs. The measurement of drug solubility is known as an important key point towards the development of the supercritical technology. If a specific drug possesses enough solubility in the solvent, its process can be feasible via the supercritical technology[9–11].

Over the last fifteen years, development of mathematical modeling through artificial intelligence (AI) and machine learning (ML) approaches have found its undeniable role on various research and development (R&D)/industrial investigations such as membrane separation, pharmaceutics, chemical reactors, nanotechnology and so on. Owing to significant cost of experimental investigation of drugs solubility in laboratory, ML techniques have paved the way to predict drugs solubility because of their brilliant advantages such as automated nature and predictive ability.[12–15].

Machine Learning (ML) is the most popular discipline for modelling data, and it may be regarded as the cornerstone of the subject of Data Science (DS). Supervised ML utilizes many approaches like regression trees, vector machines, and neural networks to train the computer. This model plays multiple applications in various scientific fields, mainly were challenging and costly experiments are performed. This branch of artificial intelligence predicts and models future data based on existing data[16–18]. Decision Trees are one of the most popular ML models. The central premise of a decision tree is to divide a complex problem into numerous more straightforward problems, which may result in a solution that is easier to grasp. Data features are predictor variables in a decision tree methodology, whereas the class to be mapped is the target variable[19].

Boosting is a common and essential strategy in ensemble learning called enhanced learning. By integrating the essential predictors, boosting enhances prediction outcomes. AdaBoost is a popular Boosting technique can add numerous base learners to provide more better estimations[20–22]. NU-SVR is another base predictor. Epsilon-SVR and NU-SVR are distinguished by the way the training problem is parametrized. Both cost functions incorporate a form of hinge loss. The nu parameter in NU-SVR allows for control over the quantity of support vectors included in the resultant model. The exact identical problem can be solved with the necessary parameters.

In this investigation, Decision Tree (DT), Adaptive Boosted Decision Trees (ADA-DT), and Nu-SVR regression models are utilized for the first time as a novel model on the available data. With an R-squared score, DT, ADA-DT, and Nu-SVR showed results of 0.836, 0.921, and 0.813, respectively. Also, in terms of MAE, they showed error rates of 4.30E−06, 1.95E−06, and 3.45E−06. Another metric is RMSE, in which DT, ADA-DT, and Nu-SVR showed error rates of 4.96E−06, 2.34E−06, and 5.26E−06, respectively. Through the analysis outputs, ADA-DT has been considered as more significant and novel model to develop and enhance the solubility of tamoxifen.

Data set. In this study, we are working with a tiny dataset that includes two inputs comprising X1 = P(bar) and X2 = T(K) (K). Also, output is Y = solubility. The number of data are 32 points retrieved from[23]. Dataset has been demonstrated below in Table 1[24].

Methodology

Decision tree. Trees are the significant data structures in various fields of artificial intelligence. A decision tree (DT) is a procedure commonly used to analyses data. A decision tree may handle either regression or classification tasks. A typical decision tree is made up of decision nodes (make a query on an input features), edges (result of a query and pass to the child node), and terminal or leaf nodes (generate the output)[25–27], as shown in Fig. 1.

Each feature of a dataset is handled as a node or hub in the DT, via the root node to be unmatched. This approach will be more developed till a leaf node is identified. The decision tree's output can be the terminal node[19,28,29]. Some of the well-known decision tree induction algorithms such as CART[19], CHAID[25], C4.5, and C5.0[27,30].

AdaBoost. Freund and Schapire invented the AdaBoost[31] to solve the binary classification problem. In AdaBoost method, the fundamental concept is to create several weak predictors sequentially using the training data subset and then merge them using a given technique. First, an equal-weighted training data is used to build the weak predictor. However, the weights of the examples in the training subset that were incorrectly estimated are raised. The new weighted training data is then used to build the weak predictor for the next round. After repeating the above technique, multiple weak predictors are obtained, and each predictor is assigned a score based on the related classification error. Using some rule to combine all weak predictors will result in a final strong predictor. Multiple AdaBoost variants have been implemented, each with its advantages and purposes[31–33].

Each x_i instance's weight w_i is set proportionally to the possibility of being accurately estimated, and implicitly proportionally to the predictor T_t error t. Furthermore, each predictor decision on a new example's final prediction is weighted according to its performance during the learning[22,34,35].

Following steps generally shows AdaBoost workflow:

- Begin with uniform sample weights.
- Initial number of predictors: M.
- For k in [1,…,M]:

No	X1 = P(bar)	X2 = T(K)	Y(solubility)
1	120	308	4×10^{-06}
2	160	308	4.94×10^{-06}
3	200	308	5.49×10^{-06}
4	240	308	5.96×10^{-06}
5	280	308	3.99×10^{-06}
6	320	308	3.88×10^{-06}
7	360	308	8.38×10^{-06}
8	400	308	1.24×10^{-05}
9	120	318	2.15×10^{-06}
10	160	318	5.79×10^{-06}
11	200	318	8.95×10^{-06}
12	240	318	7.27×10^{-06}
13	280	318	3.40×10^{-06}
14	320	318	7.03×10^{-05}
15	360	318	4.01×10^{-06}
16	400	318	1.39×10^{-05}
17	120	328	1.79×10^{-06}
18	160	328	5.13×10^{-06}
19	200	328	1.05×10^{-06}
20	240	328	5.48×10^{-05}
21	280	328	2.31×10^{-05}
22	320	328	2.04×10^{-05}
23	360	328	2.50×10^{-05}
24	400	328	4.41×10^{-05}
25	120	338	1.52×10^{-06}
26	160	338	3.84×10^{-06}
27	200	338	1.05×10^{-05}
28	240	338	2.08×10^{-05}
29	280	338	3.13×10^{-05}
30	320	338	1.95×10^{-05}
31	360	338	5.47×10^{-05}
32	400	338	6.0×10^{-05}

Table 1. Data set.

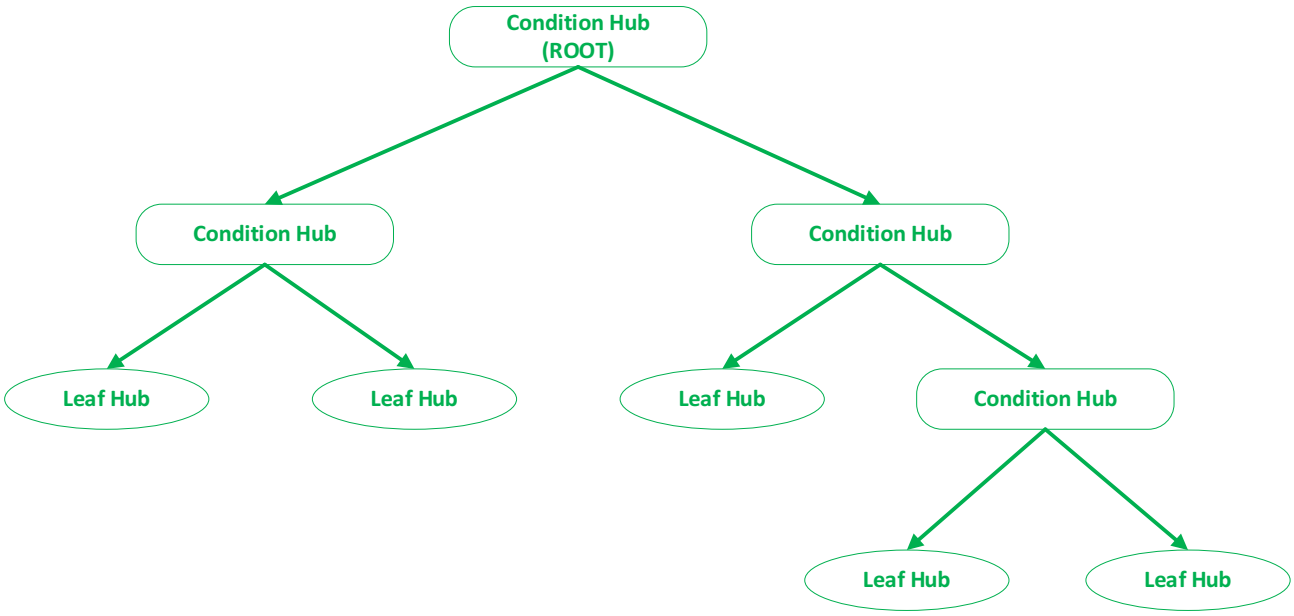

Figure 1. Decision tree sample architecture.

- Develop a base learner L_k via a weighted sample.
- Test L_k on all data.
- Set new weight for L_k using a weighted error.
- Set weights for each sample data point.

This approach has several advantages, the most prominent of which is simpler to use and requires fewer hyper-parameters to be tuned. AdaBoost is not prone to overfitting because of its design and methodology[35].

Nu-SVR. A set of input and output parameters supplied as basic configuration *{(x1, y1), …, (xn, yn)}*. The goal of the Nu-SVR method is to compute the correlation indicated in the following Equation, as *f(x)* must in neighborhood of value of *y* as possible. It should also be as flat as feasible. Since we want to avoid over-fitted models in this investigation[36–38].

$$f(x) = wT\ \Phi(x) + b \tag{1}$$

In this equation, $\Phi(x)$ is declared as the non-linear function mapping the input space to space of higher dimensions and b denotes the bias. w^T is also stands for the weight vector. Optimization is the primary objective of the task: Closeness and flatness are two of the fundamental aims of this challenge, which is why the main goal is to optimize[37–41]:

$$\frac{1}{2}|\lceil w\rceil| + \mathrm{C}\left\{ Y \cdot \varepsilon + \frac{1}{n}\sum_{i=1}^{n}\left(\xi + \xi^*\right)\right\} \tag{2}$$

According to the conditions:

$$\mathrm{y_i} - \left\langle w^T \cdot \Phi(x)\right\rangle - b \le \varepsilon + \xi_i^*, \tag{3}$$

$$\left\langle w^T \cdot \Phi(x)\right\rangle + b - y_i \le \varepsilon + \xi_i, \tag{4}$$

$$\xi_i^*, \xi_i \ge 0 \tag{5}$$

here ε is a distance between the f(x) and its actual amount. Also, ξ, ξ_i are extra slack variables depicted in[42], declares that distance of value ξ above ε error are reasonable. The parameter *C*, define as the regularization amount, indicates an equilibrium on the tolerance of error ε and flatness of *f*[38].

So, *Y* $(0 < Y < 1)$ shows the upper bound for the function of margin errors in training amounts and defines the lower bound for the fraction of support vectors. Furthermore, to address the first issue, the dual statement has been created through constructing the Lagrange function[38]:

$$\begin{aligned} \mathrm{L}: \frac{1}{2}|\lceil w\rceil|^2 + \mathrm{C}\left\{ Y \cdot \varepsilon + \frac{1}{n}\sum_{i=1}^{n}\left(\xi + \xi^*\right)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left(\eta\xi + \eta^*\xi^*\right) \\ - \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon + \xi_i + y_{i-} w^T \cdot \Phi(x) - b\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon + \xi_i + y_i + w^T \cdot \Phi(x) + b\right) - \beta\varepsilon \end{aligned} \tag{6}$$

a, a^*, η, η^*, β demonstrate the Lagrange multipliers, then $a^{(*)} = a \cdot a^*$, through maximize Lagrange function $W = \sum_{i=1}^{n}\left(a_i - a_i^*\right) \cdot \Phi(x)$ and leads to a problem with dual optimization[38]:

$$\text{Maximizes} - \frac{1}{2}\sum_{i=1}^{n}\left(a_i - a_i^*\right) \cdot \left(a_j a_j^*\right) \cdot k\left(x_i x_j\right) + \sum_{i=1}^{n} y_i\left(a_i - a_i^*\right); \tag{7}$$

Subject to:

$$\sum_{i=1}^{n}\left(a_i - a_i^*\right) = 0 (8) \tag{8}$$

$$\sum_{i=1}^{n}\left(a_i - a_i^*\right) \le CY \tag{9}$$

$$a_i, a_i^* \epsilon \left[0, \frac{C}{n}\right]. \tag{10}$$

Since $K(x_i,x_j)$ stands for the kernel function defined through $K(x_i,x_j) = \Phi(x_i)^T \cdot \Phi(x_j)$. The solution to recent Formula yields to the Lagrange multipliers a, a^*. An estimate of the function *(L)* is obtained when weight *W* is swapped into recent equations:

$$f(x) = \sum_{n=1}^{n} \left(a_i - a_i^*\right) \cdot k(x_i, x) + b \tag{11}$$

Tamoxifen targets beside estrogen receptors. Pubchem web site was used for smiles retrieval of tamoxifen (https://pubchem.ncbi.nlm.nih.gov/compound/Tamoxifen#section=InChI). Smiles code obtained was as the following (CCC(=C(C1=CC=CC=C1)C2=CC=C(C=C2)OCCN(C)C)C3=CC=CC=C3), this code was fed into LigTMap web server (https://cbbio.online/LigTMap/) to search for other molecular targets of tamoxifen, selected target classes in this search are Anticogulant, Beta_secretase, Bromodomain, Carbonic_Anhydrase, Hydrolase, Isomerase, Kinase,Ligase, Peroxisome, Transferase, Diabetes, HCV, Hpyroli, HIV, Influenza and Tuberculosis. Also, this smile code was inserted in swissADME web server (http://www.swissadme.ch/index.php) to investigate its boiled egg model in addition to the physicochemical parameters.

Results

After tuning of important hyper-parameters by run different combinations some metrics are needed to evaluate the accuracy of final models. The statistical measurements of RMSE, MAE, and R-squared is used to compare the accuracy of different models' predictions[43,44].

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{a} [z' - z]^2}{a}} \tag{12}$$

$$\mathrm{MAE} = \frac{1}{a} \sum_{j=1}^{a} |z' - z| \tag{13}$$

$$R^2 = \frac{a \sum z'z - \sum z' z}{\sqrt{\left[a \sum z'^2 - \left(\sum z'\right)^2\right]\left[a \sum z^2 - \left(\sum z\right)^2\right]}} \tag{14}$$

The above three equations are used to calculate these metrics. Here z′ and z indicates the assessed and actual data, and *a* quantity of data.

Figures 2, 3 and 4 numerically compare the real outcomes with estimated results in DT, ADA-DT and Nu-SVR machine-learning based models. As demonstrated, the ADA-DT model enjoys the greatest accuracy due to the presence of most points in a reasonable neighborhood of actual values. Considering the values of R^2 and RMSE presented in Table 2, the ADA-DT is selected as an accurate model with the best generality.

Figure 5 shows the result for assessing the influence of pressure and temperature as inputs on the solubility. Moreover, Figs. 6 and 7 illustrate two-dimensional depictions to individually analyzed the trends of two inputs on drug solubility[24]. Analysis of the figures implies the fact that increment of the pressure from 120 to 400 bar eventuates in a significant improvement in the solubility of tamoxifen. An enhancement of pressure significantly improves the amount of density accompanying with the solvating power, which positively enhances the solubility of drug in the $SCCO_2$ system. About temperature, it must be said that the results show some complexities. In detail, the modeling outcomes show that the due to the existence of a threshold pressure, the influence of

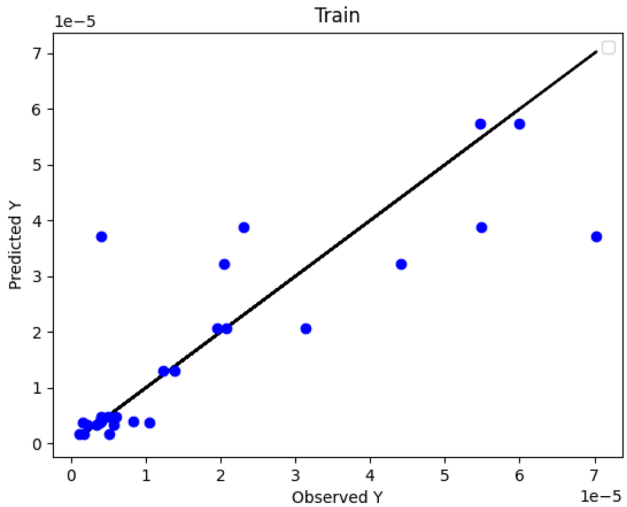

Figure 2. DT Model: actual versus predicted values/Y: solubility.

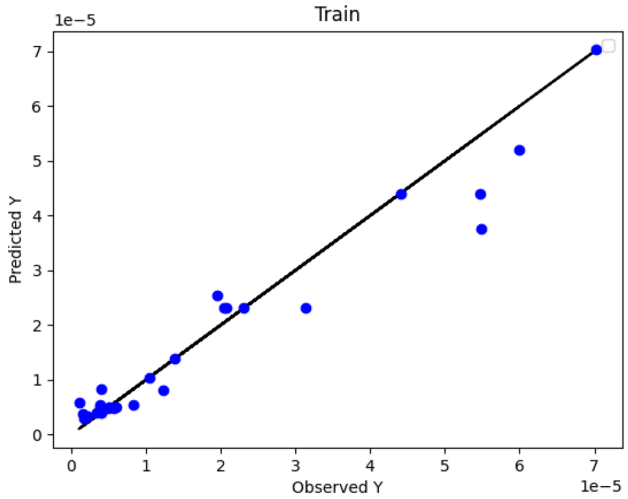

Figure 3. ADA-DT Model: actual versus predicted values/Y: solubility.

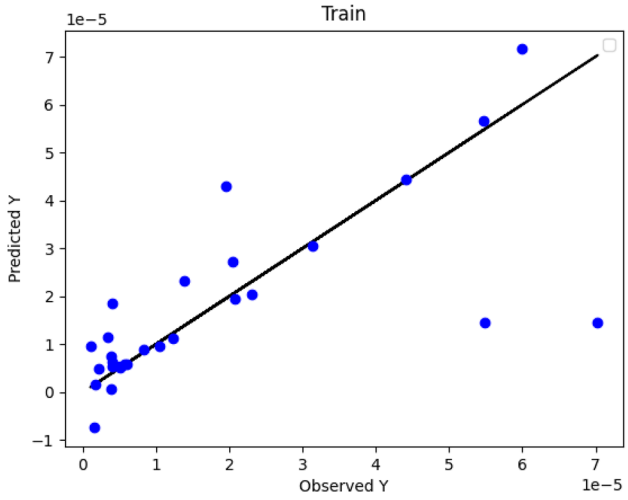

Figure 4. NU-SVR Model: actual versus predicted values/Y: solubility.

Models	MAE	RMSE	R^2
Decision tree (DT)	4.30E−06	4.96E−06	0.836
ADA-DT	1.95E−06	2.34E−06	0.921
Nu-SVR	3.45E−06	5.26E−06	0.813

Table 2. Outputs of different models.

temperature show a reversal trend. In details, at the operational pressure lower than 240 bar, increasing the temperature decreases the solubility of tamoxifen because of a decrement in the density of solvent, with negative effect on the solvating power. According to the abovementioned analysis, it is proved that a shifting pressure named cross-over pressure is existed for the values less than this pressure (lower than 240 bar), the density reduction overcomes the sublimation pressure and therefore, the solubility of tamoxifen declines[23]. When the pressure goes beyond the cross-over pressure (240 bar), the role of pressure sublimation dominates the impact of density. Thus, by increasing the pressures at the pressures higher than cross-over pressure, the solubility of tamoxifen in the $SCCO_2$ system increases[45,46]. The optimal values of pressure and temperature to obtain the highest amount of tamoxifen solubility is presented in Table 3[24].

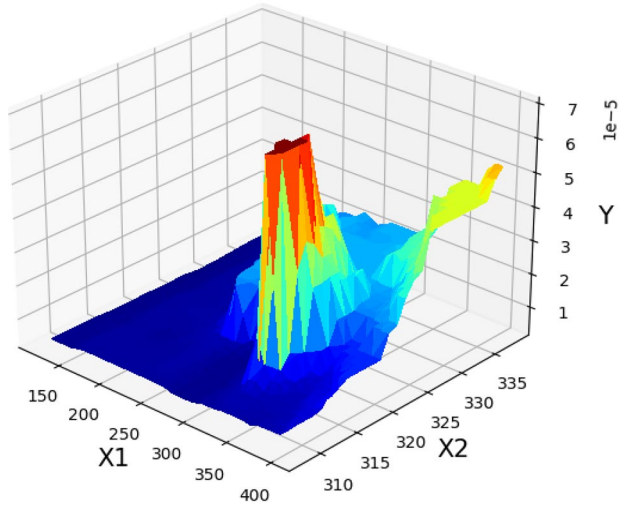

Figure 5. 3D demonstrateion of inputs/outputs/Y: solubility/X_2: temperature/X_1: pressure.

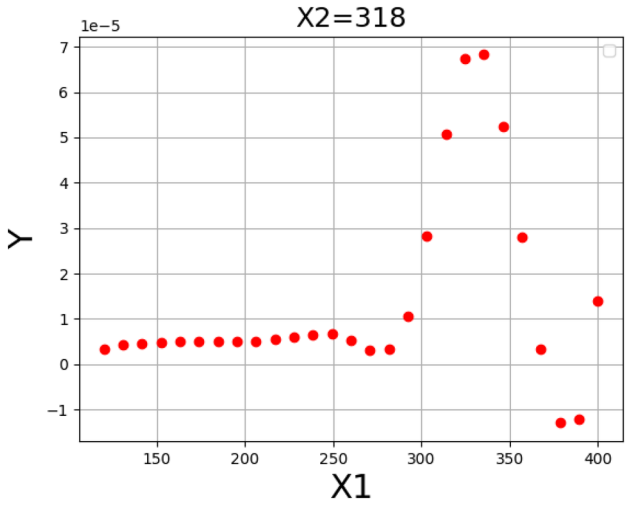

Figure 6. Solubility based on pressure/Y: solubility/X_1: pressure.

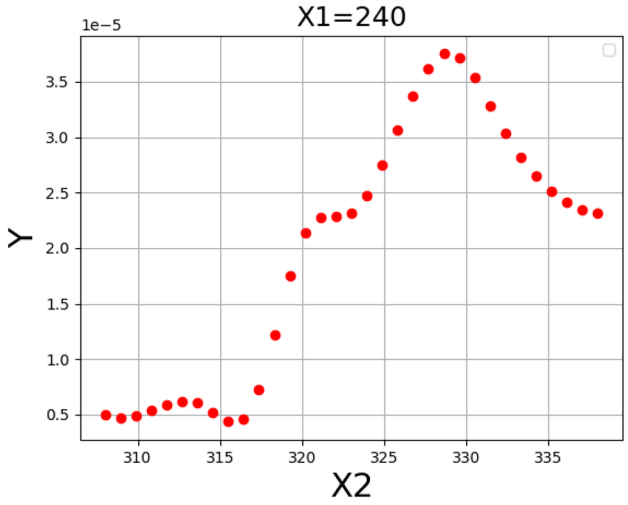

Figure 7. Solubility based on temperature/s: solubility/X_2: temperature.

X1 = P(bar)	X2 = T(K)	Y(solubility)
309	317.39	7.03e−05

Table 3. Optimized pressure and temperature/ optimized solubility.

Tamoxifen targets beside estrogen receptors and its boiled-egg model. Tamoxifen continues to be used in treatment of estrogen positive breast cancer[47]. In the current research work we decided to investigate if there are other molecular targets for this crucial drug to figure a new way in its medicinal usage. We have used CADD techniques in our previous research work[48–51] as they are useful tools in investigating diverse properties for different molecules. Through usage of SwissADME web server we could get the boiled-egg model of tamoxifen (Fig. 8 and supplementary data) that illustrates that tamoxifen with poor probability to penetrate BBB in addition to its poor GI-absorption. Additionally, the model showed that tamoxifen is PGP + which means that it can be effluated outside the cells by the action of P-glycoprotein. Being a substrate for P-glycoprotein increases the possibility of tamoxifen resistance. Improvement of tamoxifen solubility may lead to better physicochemical properties and better GI-absorbance.

Furthermore, the other possible targets for tamoxifen were explored in this research work through LigTMap web server, all disease target classes were selected except estrogen. The obtained results revealed other seven putative tamoxifen targets other than estrogen (supplementary data files), these targets are divided into three Hydrolases (CES1 protein, bifunctional epoxide hydrolase 2 and LEUKOTRIENE A-4 HYDROLASE), two HCV (NON-STRUCTURAL PROTEIN 4A, SERINE PROTEASE NS3 and RNA-directed RNA polymerase), one predicted protein target for Beta_secretase (BETA-SECRETASE 1) and one protein target for Bromodomain (Bromodomain-containing protein 4). These plausible targets for tamoxifen are ranked according to LigTMap score as shown in Table 4, tamoxifen showed ligand similarity for these targets with range from 40 to 69%, the best ligand similarity score (0.689) was assigned for CTX ligand in CES1 protein (pdb Id: 1ya4). The results also revealed more than 55% binding similarity with Y80 lignad in Bromodomain-containing protein 4 (pdb ID: 4yh3). The best docking score was -7.925 kcal/mol with CES1 protein (pdb ID: 1ya4). Additionally, tamoxifen showed good docking score energy with these seven putative targets as shown in Table 4, docking score ranged from -5.759 to -7.925 kcal/mol. Figure 9 represents the 2D interactions of tamoxifen with CES1 protein binding site.

Conclusion

In this research, to predict tamoxifen solubility, supercritical carbon dioxide is used as solvent. Experimental data have been provided through the literature, then analyzed to develop a predictive model. On the provided data, Decision Tree (DT), Adaptive Boosted Decision Trees (ADA-DT), and Nu-SVR regression models are employed

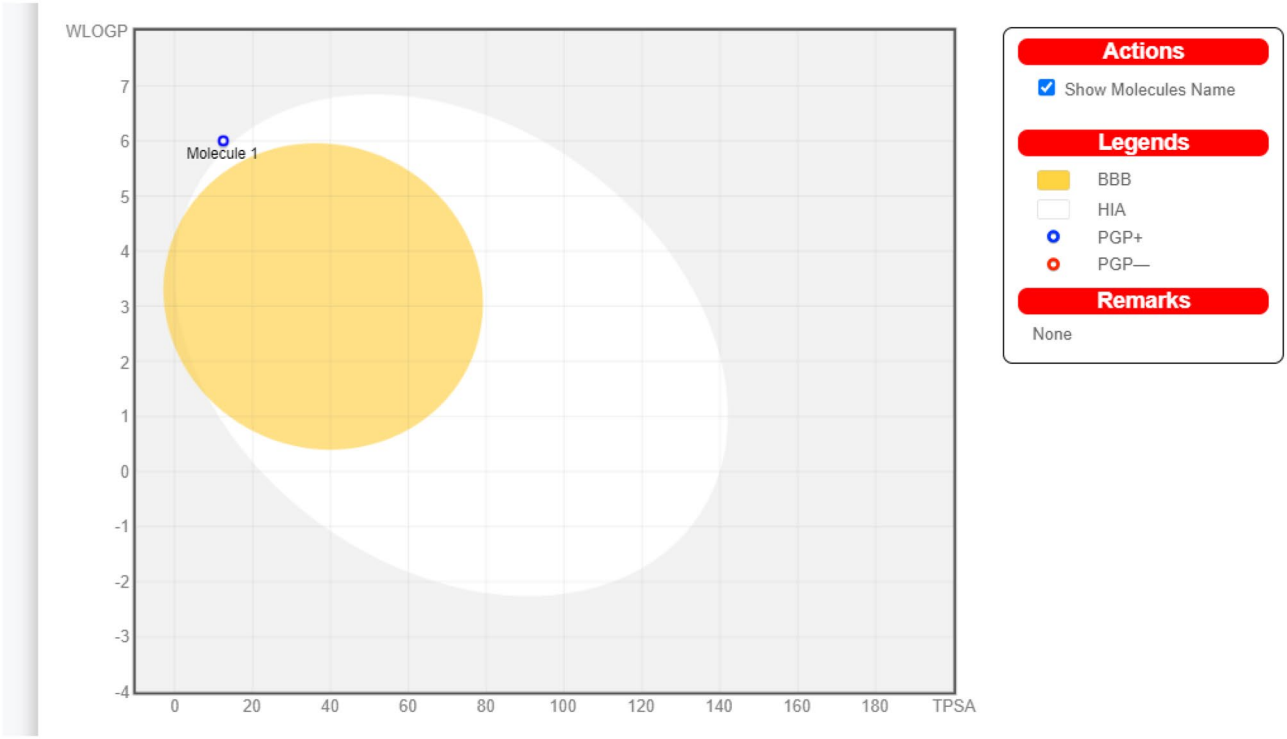

Figure 8. Tamoxifen boiled-egg model by SwissADME.

Rank	Target class	Pdb ID	PSOVina2 docking score (kcal/mol)
1	Hydrolase	1ya4	−7.925
2	Bromodomain	4yh3	−6.148
3	HCV	4b6f	−6.634
4	Hydrolase	4y2t	−7.418
5	Hydrolase	5aen	−7.729
6	Beta_secretase	1w51	−5.759
7	HCV	3lkh	−6.717

Table 4. Predicted putative tamoxifen targets retrieved from LigTMap.

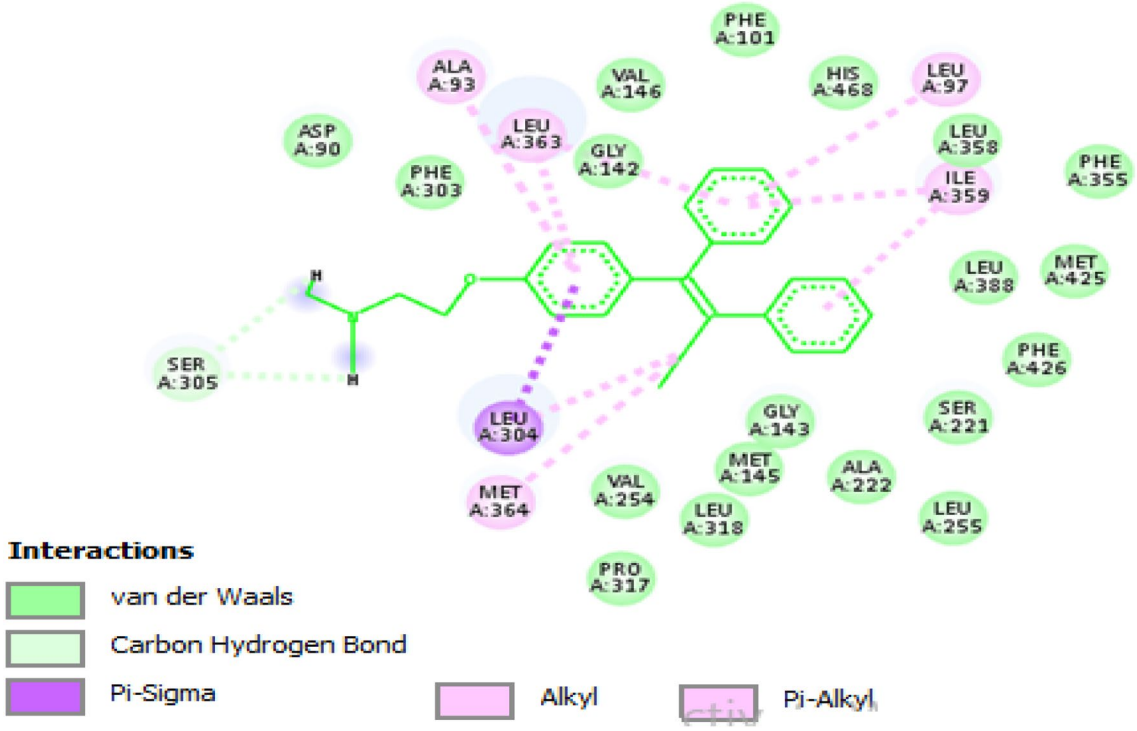

Figure 9. 2D interactions of tamoxifen with CES1 protein binding site (pdb ID: 1ya4).

through two parameters as inputs, Pressure and Temperature. Furthermore, solubility considered as output. DT, ADA-DT, and Nu-SVR demonstrate R-squared scores of 0.836, 0.921, and 0.813. The MAE error has been demonstrated by the rates of 4.30E−06, 1.95E−06, and 3.45E−06. RMSE as another statistic, revealed the error rates of 4.96E−06, 2.34E−06, and 5.26E−06 for DT, ADA-DT, and Nu-SVR, respectively. Based on these measurements and some visual inspection, ADA-DT has been considered as the best model to identify optimal values to predict drug solubility based on the optimized values x1 = 309, x2 = 317.39, Y1 = 7.03e−05). Furthermore, LigTMap web server has helped in identification of seven putative tamoxifen protein targets other than estrogen.

Data availability

All data are available within the published paper.

Received: 22 May 2022; Accepted: 1 December 2022
Published online: 24 January 2023

References

1. Atkinson, A. J. Chapter 1—Introduction to clinical pharmacology. In *Atkinson's Principles of Clinical Pharmacology* 4th edn (eds Huang, S.-M. *et al.*) 1–10 (Academic Press, 2022).
2. Zhuang, W. *et al.* Ionic liquids in pharmaceutical industry: A systematic review on applications and future perspectives. *J. Mol. Liq.* **349**, 118145 (2022).
3. Mohs, R. C. & Greig, N. H. Drug discovery and development: Role of basic biological research. *Alzheimer's Dementia Transl. Res. Clin. Interv.* **3**(4), 651–657 (2017).
4. Berdigaliyev, N. & Aljofan, M. An overview of drug discovery and development. *Future Med. Chem.* **12**(10), 939–947 (2020).

5. Elveny, M. *et al.* A state-of-the-art review on the application of various pharmaceutical nanoparticles as a promising technology in cancer treatment. *Arab. J. Chem.* **14**(10), 103352 (2021).
6. Padrela, L. *et al.* Supercritical carbon dioxide-based technologies for the production of drug nanoparticles/nanocrystals—A comprehensive review. *Adv. Drug Deliv. Rev.* **131**, 22–78 (2018).
7. De Zordi, N. *et al.* Solubility of pharmaceutical compounds in supercritical carbon dioxide. *J. Supercrit. Fluids* **66**, 16–22 (2012).
8. Pishnamazi, M. *et al.* Evaluation of supercritical technology for the preparation of nanomedicine: Etoricoxib analysis. *Chem. Eng. Technol.* **44**(3), 559–564 (2021).
9. Khoshmaram, A. *et al.* Supercritical process for preparation of nanomedicine: Oxaprozin case study. *Chem. Eng. Technol.* **44**(2), 208–212 (2021).
10. Khaw, K.-Y. *et al.* Solvent supercritical fluid technologies to extract bioactive compounds from natural sources: A review. *Molecules* **22**(7), 1186 (2017).
11. Darani, K. K. & Mozafari, M. R. Supercritical fluids technology in bioprocess industries: A review. *J. Biochem. Technol.* **2**(1), 144–152 (2010).
12. Babanezhad, M. *et al.* Pattern recognition of the fluid flow in a 3D domain by combination of Lattice Boltzmann and ANFIS methods. *Sci. Rep.* **10**(1), 1–13 (2020).
13. Nguyen, Q. *et al.* Prediction of thermal distribution and fluid flow in the domain with multi-solid structures using cubic-interpolated pseudo-particle model. *PLoS ONE* **15**(6), e0233850 (2020).
14. Babanezhad, M. *et al.* Influence of number of membership functions on prediction of membrane systems using adaptive network based fuzzy inference system (ANFIS). *Sci. Rep.* **10**(1), 1–20 (2020).
15. Nguyen, Q. *et al.* Thermal and flow visualization of a square heat source in a nanofluid material with a cubic-interpolated pseudo-particle. *ACS Omega* **5**(28), 17658–17663 (2020).
16. Carbonell, J. G., Michalski, R. S. & Mitchell, T. M. An overview of machine learning. *Mach. Learn.* 3–23 (1983).
17. Mitchell, T. M. *The Discipline of Machine Learning* Vol. 9 (Carnegie Mellon University, School of Computer Science, Machine Learning, 2006).
18. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
19. Breiman, L. *et al. Classification and Regression Trees* (Routledge, 2017).
20. Bartlett, P. *et al.* Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Stat.* **26**(5), 1651–1686 (1998).
21. Feng, D.-C. *et al.* Failure mode classification and bearing capacity prediction for reinforced concrete columns based on ensemble machine learning algorithm. *Adv. Eng. Inform.* **45**, 101126 (2020).
22. Ying, C. *et al.* Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica* **39**(6), 745–758 (2013).
23. Pishnamazi, M. *et al.* Thermodynamic modelling and experimental validation of pharmaceutical solubility in supercritical solvent. *J. Mol. Liq.* **319**, 114120 (2020).
24. 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* **27**(16), 5140 (2022).
25. Quinlan, J. R. Learning decision tree classifiers. *ACM Comput. Surv. (CSUR)* **28**(1), 71–72 (1996).
26. Xu, M. *et al.* Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.* **97**(3), 322–336 (2005).
27. Namazi, N. I. *et al.* Solubility enhancement of decitabine as anticancer drug via green chemistry solvent: Novel computational prediction and optimization. *Arab. J. Chem.* **15**(12), 104259 (2022).
28. Kushwah, J. S. *et al.* Comparative study of regressor and classifier with decision tree using modern tools. In *Materials Today Proceedings* (2021).
29. Mathuria, M. Decision tree analysis on j48 algorithm for data mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**(6) (2013).
30. Segal, M. R. & Bloch, D. A. A comparison of estimated proportional hazards models and regression trees. *Stat. Med.* **8**(5), 539–550 (1989).
31. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**(1), 119–139 (1997).
32. Schapire, R. E. & Singer, Y. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* **37**(3), 297–336 (1999).
33. Webb, G. I. Multiboosting: A technique for combining boosting and wagging. *Mach. Learn.* **40**(2), 159–196 (2000).
34. Cao, J., Kwong, S. & Wang, R. A noise-detection based AdaBoost algorithm for mislabeled data. *Pattern Recogn.* **45**(12), 4451–4465 (2012).
35. Krithiga, R. & Ilavarasan, E. Hyperparameter tuning of AdaBoost algorithm for social spammer identification. *Int. J. Pervasive Comput. Commun.* (2021).
36. Bhatt, D. *et al.* An enhanced mems error modeling approach based on nu-support vector regression. *Sensors* **12**(7), 9448–9466 (2012).
37. Moosaei, H. *et al.* Generalized twin support vector machines. *Neural Process. Lett.* **53**(2), 1545–1564 (2021).
38. Zhao, Z. *et al.* Multi support vector models to estimate solubility of Busulfan drug in supercritical carbon dioxide. *J. Mol. Liq.* **350**, 118573 (2022).
39. Meyer, D., Leisch, F. & Hornik, K. The support vector machine under test. *Neurocomputing* **55**(1–2), 169–186 (2003).
40. Ralaivola, L. & d'Alché-Buc, F. Incremental support vector machine learning: A local approach. In *International Conference on Artificial Neural Networks*. (Springer, 2001).
41. Rodriguez-Galiano, V. *et al.* Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **71**, 804–818 (2015).
42. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
43. Garosi, Y. *et al.* Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. *Sci. Total Environ.* **664**, 1117–1132 (2019).
44. Pourghasemi, H. R. *et al.* Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process. *Geosci. Front.* **11**(6), 2207–2219 (2020).
45. Pishnamazi, M. *et al.* Using static method to measure tolmetin solubility at different pressures and temperatures in supercritical carbon dioxide. *Sci. Rep.* **10**(1), 1–7 (2020).
46. Zhu, H. *et al.* Machine learning based simulation of an anti-cancer drug (busulfan) solubility in supercritical carbon dioxide: ANFIS model and experimental validation. *J. Mol. Liq.* **338**, 116731 (2021).
47. Abdellatif, K. R. A., Belal, A. & Omar, H. A. Design, synthesis and biological evaluation of novel triaryl (Z)-olefins as tamoxifen analogues. *Bioorg. Med. Chem. Lett.* **23**(17), 4960–4963 (2013).
48. Belal, A. 3D-pharmacophore modeling, molecular docking, and virtual screening for discovery of novel CDK4/6 selective inhibitors. *Russ. J. Bioorg. Chem.* **47**(1), 317–333 (2021).
49. Zhaorigetu, et al., Antiproliferative, apoptotic effects and suppression of oxidative stress of quercetin against induced toxicity in lung cancer cells of rats: In vitro and in vivo study. *J. Cancer* **12**(17), 5249–5259 (2021).
50. Mehany, A. B. M. *et al.* Apoptotic and anti-angiogenic effects of propolis against human bladder cancer: Molecular docking and in vitro screening. *Biomarkers* **27**(2), 138–150 (2022).
51. Belal, A. Pyrrolizines as potential anticancer agents: design, synthesis, caspase-3 activation and micronucleus (MN) induction. *Anticancer Agents Med. Chem.* **18**(15), 2124–2130 (2018).

Acknowledgements
This study is supported via funding from Prince sattam bin Abdulaziz University project number (PSAU/2023/R/1444). The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant code: (23UQU4290565DSR060). The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP2/50/43).

Author contributions
S.M.A.: supervision, modeling, validation, writing; A.S.A.: methodology, writing, investigation; M.M.A.: writing, methodology, data analysis; A.B.: supervision, writing, modeling, validation, data analysis, methodology; M.A.S.A.: writing, modeling, validation, data analysis, methodology; A.A.S.: writing, methodology, data analysis, editing, investigation; B.K.A.: writing, methodology, data analysis, editing; K.V.: writing, methodology, data analysis, editing; A.M.A.: writing, methodology, data analysis, editing, software; M.P.: supervision, writing, modeling, validation, data analysis, methodology.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-25562-y.

Correspondence and requests for materials should be addressed to S.M.A., A.B., A.M.A. or M.P.

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.