scientific reports

Check for updates

OPEN

Development a novel robust method to enhance the solubility of Oxaprozin as nonsteroidal anti-inflammatory drug based on machine-learning

Walid Kamal Abdelbasset[1,2]✉, Safaa M. Elkholi[3], Khadiga Ahmed Ismail[4], Sameer Alshehri[5], Ahmed Alobaida[6], Bader Huwaimel[7], Ahmed D. Alatawi[8], Amal M. Alsubaiyel[9]✉, Kumar Venkatesan[10] & Mohammed A. S. Abourehab[11,12]✉

Accurate specification of the drugs' solubility is known as an important activity to appropriately manage the supercritical impregnation process. Over the last decades, the application of supercritical fluids (SCFs), mainly CO_2, has found great interest as a promising solution to dominate the limitations of traditional methods including high toxicity, difficulty of control, high expense and low stability. Oxaprozin is an efficient off-patent nonsteroidal anti-inflammatory drug (NSAID), which is being extensively used for the pain management of patients suffering from chronic musculoskeletal disorders such as rheumatoid arthritis. In this paper, the prominent purpose of the authors is to predict and consequently optimize the solubility of Oxaprozin inside the CO_2SCF. To do this, the authors employed two basic models and improved them with the Adaboost ensemble method. The base models include Gaussian process regression (GPR) and decision tree (DT). We optimized and evaluated the hyper-parameters of them using standard metrics. Boosted DT has an MAE error rate, an R2-score, and an MAPE of 6.806E-05, 0.980, and 4.511E-01, respectively. Also, boosted GPR has an R2-score of 0.998 and its MAPE error is 3.929E-02, and with MAE it has an error rate of 5.024E-06. So, boosted GPR was chosen as the best model, and the best values were: (T = 3.38E + 02, P = 4.0E + 02, Solubility = 0.001241).

Recently, numerous endeavors have been made to find various efficacious and promising techniques to appropriately solve or at least mitigate the unprecedented challenges of pharmaceutical industry such as low solubility of drugs, unacceptable productivity and growing research and development (R&D) costs[1–3]. Oxaprozin (known under the brand name Daypro) is a well-known nonsteroidal anti-inflammatory drug (NSAID), which has indication in pain/swelling management of adult patients suffering from rheumatoid arthritis, ankylosing

[1]Department of Health and Rehabilitation Sciences, College of Applied Medical Sciences, Prince Sattam Bin Abdulaziz University, P.O. Box. 173, Al-Kharj 11942, Saudi Arabia. [2]Department of Physical Therapy, Kasr Al-Aini Hospital, Cairo University, Giza 12613, Egypt. [3]Department of Rehabilitation Sciences, College of Health and Rehabilitation Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box. 84428, 11671 Riyadh, Saudi Arabia. [4]Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Taif University, P.O.BOX.11099, Taif 21944, Saudi Arabia. [5]Department of Pharmaceutics and Industrial Pharmacy, College of Pharmacy, Taif University, P.O.Box 11099, Taif 21944, Saudi Arabia. [6]Department of Pharmaceutics, College of Pharmacy, University of Hail, Hail 81442, Saudi Arabia. [7]Department of Pharmaceutical Chemistry, College of Pharmacy, University of Hail, Hail 81442, Saudi Arabia. [8]Department of Clinical Pharmacy, College of Pharmacy, Jouf University, Sakaka, Al-Jouf, Saudi Arabia. [9]Department of Pharmaceutics, College of Pharmacy, Qassim University, Buraidah 52571, Saudi Arabia. [10]Department of Pharmaceutical Chemistry, College of Pharmacy, King Khalid University, Abha 62529, Kingdom of Saudi Arabia. [11]Department of Pharmaceutics, College of Pharmacy, Umm Al-Qura University, Makkah 21955, Saudi Arabia. [12]Department of Pharmaceutics and Industrial Pharmacy, Faculty of Pharmacy, Minia University, Minia 61519, Egypt. ✉email: walidkamal.wr@gmail.com; asbiel@qu.edu.sa; maabourehab@uqu.edu.sa

spondylitis and soft tissue disorders[4]. Microsomal oxidation and glucuronic acid conjugation are known as the major procedures of Oxaprozin primary metabolism in the liver. Metabolism of this drug in the liver results in the formation of Ester and ether glucuronides as the prominent conjugated metabolites. Manageable safety profile, great efficiency, low liver toxicity and appropriate cost has made Oxaprozin a golden NSAID for the pain alleviation of patients with chronic musculoskeletal diseases[5–7].

Application of novel approaches to increase the poor solubility of drugs is an attractive approach to solve one of the challenges of pharmaceutical industry. Recently, the use of supercritical fluids (SCFs) for processing therapeutic agents has offered suitable opportunities for the pharmaceutical manufacturing scientists[8]. This type of fluid possesses great potential of application in disparate scientific scopes including drug delivery, chromatography, and extraction[9]. Among various sorts of SCFs, supercritical carbon dioxide (SCCO$_2$) recommends various interesting technological advantages such as low toxicity, ignorable flammability and environmentally friendly characteristic which may eventuate in result a significant decrement in the application of commonly employed organic solvents. Apart from different industrial-based applications, particle micronization using SCCO$_2$ is one of the novel and promising approaches for fabricating micro-/nanoparticles with controlled size and purity[10].

Prediction of drugs solubility using artificial intelligence (AI) method has currently attracted the attention as a noteworthy option for validating the actual data obtained from experimental research. Development of predictive modeling and simulation via this technique for different industries (i.e., separation, purification, extraction and drug delivery) can considerably decline computation time and guarantee the accuracy of conducted experimental results[11,12].

Computers can learn from data without having to be explicitly programmed, using a class of AI techniques known as machine learning (ML). Machine learning seeks to develop meta-programs that process experimentally gathered data and apply it to train models for the prediction of unknown future inputs[13,14]. Ensemble methods are also a class of ML methods that use several basic models to achieve higher accuracy and generality in prediction[15,16].

When multiple weak estimators are combined to produce a robust estimator, it is known as "boosting." Because of the sequential logic employed by Boosting, each weak estimator has a direct impact on its successor. Particularly AdaBoost[17] is a typical boosting algorithm that uses reweighted training data to gradually obtain weak classifiers. It was decided to use Adaboost procedures to modify the efficiency of two base estimators as the foundation of this study. Decision Tree and Gaussian process regression are selected base models.

Decision Tree asks a series of questions using feature sets, such as 'is equal' or 'is greater,' and based on the provided answers, another question is asked to respond. Same procedure is repeated until no further inquiries are received, at the point the result is obtained. The data is constantly divided into binary components, allowing the Decision Tree to grow. To evaluate the divisions for all attributes, a randomness metric such as entropy is used[18,19].

Also, for both exploration and exploitation, Gaussian process regression is a non-parametric Bayesian modeling technique. The primary profitability of the method is the ability to forge a reliable response for input variables. It can describe a broad range of interactions between features and targets by using a feasibly infinite count of input features and allowing the data to define the complexity level through Bayesian inference[20,21].

Experimental

In this paper, validation of predictive models' results is done by their comparison with obtained experimental data from the experiments of Khoshmaram et al.[22]. They developed a pressure-volume-temperature (PVT) cell to experimentally measure the solubility value of Oxaprozin in SCCO$_2$ solvent[22]. In their developed setup, first, the SCCO$_2$ solvent is prepared via increasing the pressure of gaseous CO$_2$ through the liquefaction unit. In the second step, the impurities of condensed manufactured SCOO$_2$ are removed via an inline filter. Then after, the purified SCOO$_2$ flows through a surge tank before its entrance to the PVT cell. The controlling process of temperature as an important parameter directly affects the solubility value of drug takes place using heating elements that wrapped the chamber and are isolated via PTFE layer.

Data Set

The dataset used in this study comes from[22], which has just 32 data points. The temperature and pressure are two input parameters. Each vector also has one output (solubility). Table 1 shows the dataset.

Methodology

GPR.

Gaussian process regression is one of the base models used. GPR, unlike other regression models, does not necessitate the specification of an exact fitting function. A multidimensional Gaussian distribution sampled at random points can be compared to field data [23,24].

The target y is simulated as $f(\boldsymbol{x})$ for a collection of n-dimensional instances $D = \{(\boldsymbol{x}_i, y_i)|i = 1, \ldots, n\}$, where $\boldsymbol{x}_i \in R^d$ is input data point and $y_i \in R$ is the output vector.

$$y = f(\boldsymbol{x}) \tag{1}$$

The GP is declared using *f(x)*, which is an implicit function illustrated as a collection of random variables:

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), \mathbf{K}) \tag{2}$$

In the above equation, *K* denotes any covariance defined by kernels and their corresponding input values and *m(x)* is the mean operator.

No	Temperature (K)	Pressure (bar)	Solubility (mole fraction)
1	3.08E + 02	1.20E + 02	8.19E-05
2	3.08E + 02	1.60E + 02	1.58E-04
3	3.08E + 02	2.00E + 02	2.24E-04
4	3.08E + 02	2.40E + 02	2.80E-04
5	3.08E + 02	2.80E + 02	3.44E-04
6	3.08E + 02	3.20E + 02	4.06E-04
7	3.08E + 02	3.60E + 02	4.73E-04
8	3.08E + 02	4.00E + 02	5.33E-04
9	3.18E + 02	1.20E + 02	7.55E-05
10	3.18E + 02	1.60E + 02	1.41E-04
11	3.18E + 02	2.00E + 02	2.45E-04
12	3.18E + 02	2.40E + 02	3.56E-04
13	3.18E + 02	2.80E + 02	4.64E-04
14	3.18E + 02	3.20E + 02	5.58E-04
15	3.18E + 02	3.60E + 02	6.60E-04
16	3.18E + 02	4.00E + 02	7.66E-04
17	3.28E + 02	1.20E + 02	5.34E-05
18	3.28E + 02	1.60E + 02	1.28E-04
19	3.28E + 02	2.00E + 02	3.02E-04
20	3.28E + 02	2.40E + 02	4.14E-04
21	3.28E + 02	2.80E + 02	5.82E-04
22	3.28E + 02	3.20E + 02	7.87E-04
23	3.28E + 02	3.60E + 02	8.51E-04
24	3.28E + 02	4.00E + 02	1.03E-03
25	3.38E + 02	1.20E + 02	3.31E-05
26	3.38E + 02	1.60E + 02	9.09E-05
27	3.38E + 02	2.00E + 02	2.98E-04
28	3.38E + 02	2.40E + 02	4.81E-04
29	3.38E + 02	2.80E + 02	6.77E-04
30	3.38E + 02	3.20E + 02	8.89E-04
31	3.38E + 02	3.60E + 02	1.08E-03
32	3.38E + 02	4.00E + 02	1.24E-03

Table 1. The Whole Dataset.

Decision Tree.

Trees are a fundamental data structure in a variety of AI contexts. An ML technique known as decision trees (DTs) is normally usage to measure the data. It is possible to utilize a decision tree to solve different estimation issues. To build a basic decision tree, you need internal nodes (which makes decision with query input features), edges (which return results and transmit them to children), and terminal or leaf nodes (which return results and send them to children) (that make decision on final output)[25,26].

The root node is a special and unique node in the DT, which treats each dataset feature as a hub or node. To demonstrate how the tree model works, we start with a single node and work our way down the tree (output). Until a terminal node is found, this strategy will be tweaked and refined. The DT's forecast or outcome would be the terminal node[18,27,28]. The most useful algorithms for decision tree induction are CART[28], CHAID[25], C4.5, and C5.0[29].

ADABOOST.

Multiple base predictors can be combined to create an ensemble learning-based model, which outperforms a single predictor. By altering the weight distribution of samples, Freund and Schapire[17] proposed the AdaBoost algorithm for enhancing the accuracy of weak learners. Because of its advantages, this method has become increasingly popular[30,31].

As the "AdaBoost" name implies, this technology adaptively enhances base predictors, enabling them to address complicated issues. One of the symptom for theamicability of basic models is that they have good generalization properties due to their simple structure. But despite the fact that they are easy to use in real-world situations, their architecture is severely biased, therefore they cannot handle complex jobs.

The Adaboost algorithm from Hastie et al.[32,33] is mostly demonstrated in the following steps.

1. Set weights for data points:

$$\omega_i = \frac{1}{N}, i \in \{1, \ldots, N\} \tag{3}$$

2. Set Number of base estimators as M.
3. For b from 1 to M:

(a) Develop a learner $G_b(x)$ using the weights ω_i.

$$\text{(b)}\quad err_b = \frac{\sum_{i=1}^{N} \omega_i I\left(y_i \neq G_b(x_i)\right)}{\sum_{i=1}^{n} \omega_i} \tag{4}$$

$$\text{(c)}\quad \alpha_b = \log\left(\frac{1 - err_b}{err_b}\right) \tag{5}$$

$$\text{(d)}\quad \omega_i \leftarrow \omega_i . \exp\left(\alpha_b . I\left(y_i \neq G_b(x_i)\right)\right), i = 1, \ldots, N \tag{6}$$

4. Final Output:

$$\mathrm{G(x)} = sign\left(\sum_{b=1}^{M} \alpha_b G_b(x)\right) \tag{7}$$

In the previous procedure, the quantity of data vectors and the number of iterations are *N* and *M*, respectively. The estimator that passes *b* over the data is $G_b(x)$. Building a prediction model (Base model) can be done in a variety of ways, but the most frequent is to employ stumps or very short trees. The operator I is set to 0 if the logical correlation is false and to 1 if the correlation is true, as shown by the indicator variable[34–36].

Results

Important hyper-parameters of selected models were first tuned applying the search grid method to assess the efficacy of the approaches described in this study. The resultant models were then examined using three distinct criteria, as specified below: MAE, MAPE, and R-score[37,38]:

$$\text{MAE ERROR} = \tfrac{1}{\mathrm{n}} \times \sum_{i=1}^{\mathrm{n}} \left|\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right| \tag{8}$$

$$\text{MAPE ERROR} = \frac{1}{\mathrm{n}} \times \sum_{i=1}^{\mathrm{n}} \left|\frac{\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}}{\mathrm{y}_{\mathrm{i}}}\right| \tag{9}$$

The third regression performance metric in our research is R^2 score. The R^2-Score is used on a regression line to determine how close the estimated amounts are to the true (expected) amounts.

$$\mathrm{R}^2 - score = 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{i}})^2}{\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{y_i} - \upmu)^2} \tag{10}$$

μ indicates the mean of the expected data[39].

In Figs. 1 and 2, the ADA + DT and ADA + GPR models are analyzed in terms of expected values and estimated values, respectively. The blue dots are the estimated values with the training samples and the red dots with the test data. The distance from the expected data line is important to us. Also, the numerical results of the three criteria mentioned above are explained in Table 2. Based on results, the ADA + GPR model has passed almost all the points of the training data. But despite this fact, we can say that the obtained model has no overfitting problem because the red dots, which are test data and have not been included in the training phase, are also close to the expected values.

Figure 3 shows the simultaneous impact of the pressure and temperature as inputs the only output (Oxaprozin solubility). This diagram shows that increasing both inputs generally increase the output value. By keeping each of the two input parameters constant and changing the other parameter, we obtained two-dimensional Figs. 4 and 5, which confirms this fact. Figure 4 illustrates the influence of pressure and Fig. 5 demonstrates the impact of temperature on the solubility value of Oxaprozin. To analyze the diagrams, the effects of pressure and temperature on the solubility of drug must be considered. It is conspicuous from the graphs that whenever the temperature value improves, the molecular compaction in the $SCCO_2$ system increases, which consequently eventuates in enhancing the solvating power of solvent and thus, increasing the solubility of Oxaprozin[40]. Figure 4 proves nearly 8 times enhancement in the solubility value of Oxaprozin by enhancing the pressure from 110 to 410 bar.

About temperature, the presence of opposite impacts on two competing parameters makes the analysis difficult. Increasing the temperature, the sublimation pressure of $SCCO_2$ system increases that positively encourages the Oxaprozin solubility. On the other hand, increase in temperature deteriorates the density of solvent that results in reducing the solubility of drug. To evaluate simultaneous impact of these parameters, cross-over

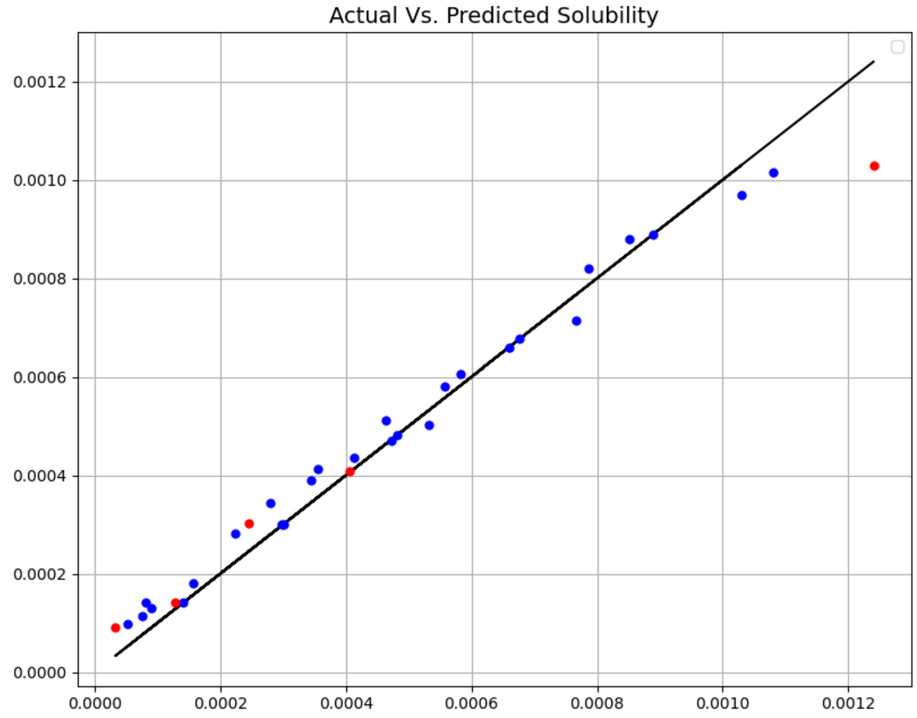

Figure 1. Expected and estimated values (ADA + DT).

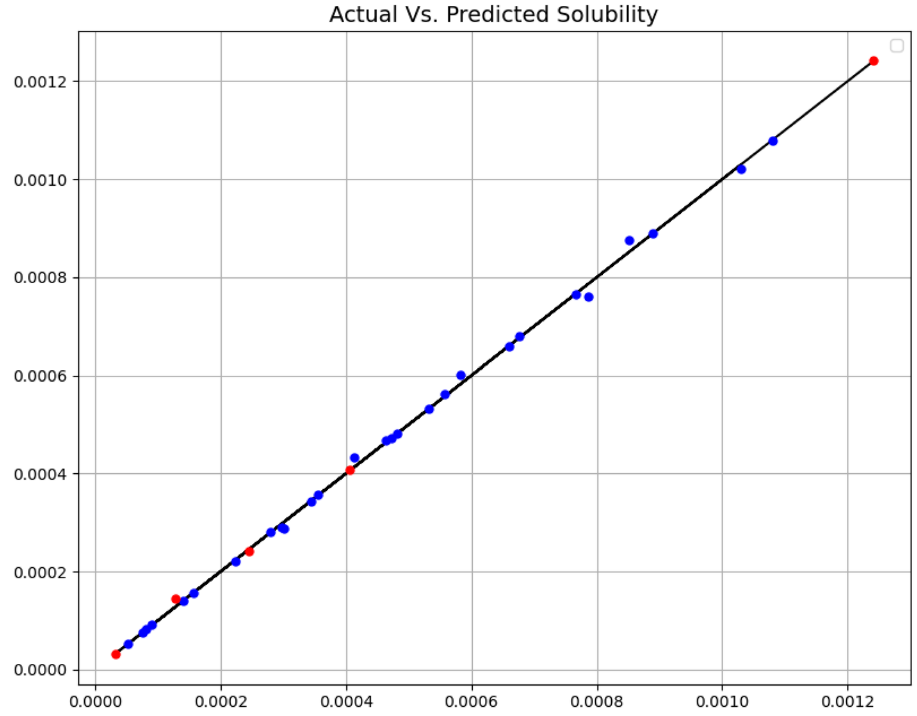

Figure 2. Expected and estimated values (ADA + GPR).

Models	MAE	R^2	MAPE
ADA + DT	6.806E-05	0.980	4.511E-01
ADA + GPR	5.024E-06	0.998	3.929E-02

Table 2. Final Model Results.

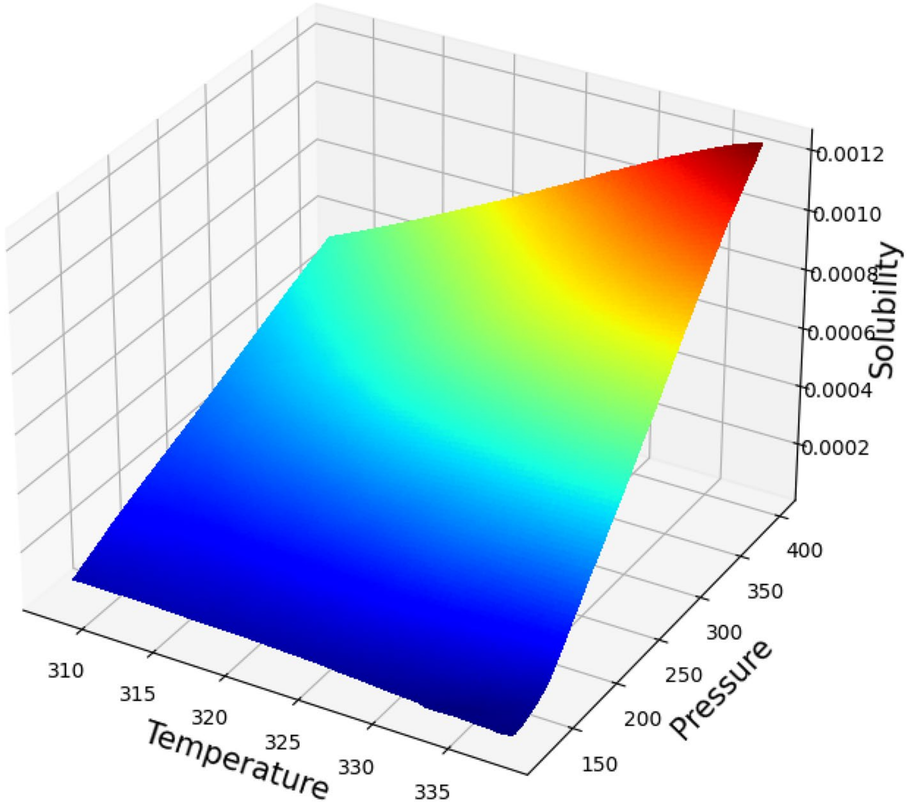

Figure 3. prediction surface in final ADA + GPR.

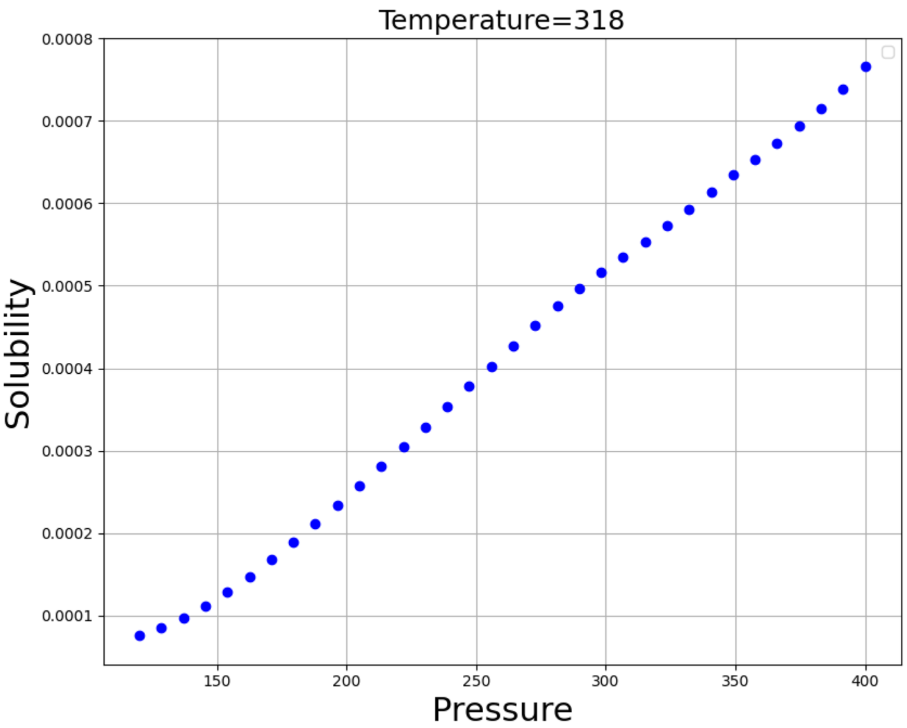

Figure 4. Trends for Pressure.

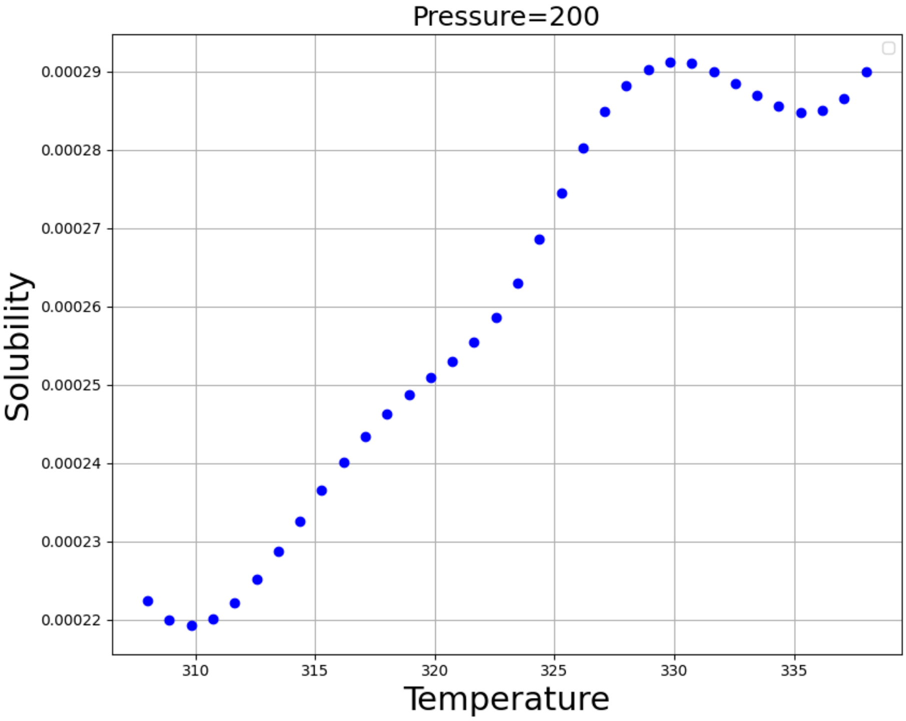

Figure 5. Trends for Temperature.

Temperature (K)	Pressure (bar)	Solubility
3.38E + 02	4. 0E + 02	0.001241

Table 3. Optimal Values.

pressure (CP) must be considered. At pressure values lower than CP, density reduction possesses stronger effect than sublimation pressure increases and therefore, when the temperature increases the solubility of Oxaprozin in $SCCO_2$ fluid reduces. At pressure values greater than CP, sublimation pressure increment has greater impact than density reduction and therefore, when the temperature increases the solubility of Oxaprozin in $SCCO_2$ fluid considerably improves. This analysis agrees with similar papers[10]. The optimal values, which should therefore be approximately the upper limit of both inputs, are also shown in Table 3, which are the same as the maximum values.

Conclusion

In current years, increasing the solubility values of different commonly employed drugs using green solvents is an attractive field of study in pharmaceutics. $SCCO_2$ has been recently introduced as a promising alternative for organic solvents because of having valuable features such as high efficacy, inflammability, and low toxicity. In this study, two base models (weak estimators) were used and boosted with Adaboost methods with the aim accurate prediction of Oxaprozin solubility in $SCCO_2$ system. Decision tree (DT) and Gaussian process regression are two of these models (GPR). We optimized these models' hyperparameters and evaluated them using standard metrics. The MAE error rate, R^2-score, and MAPE of boosted DT are 6.806E-05, 0.980, and 4.511E-01, respectively. Furthermore, boosted GPR has an R^2-score of 0.998, MAPE error of 3.929E-02, and MAE error rate of 5.024E-06. As a result, ADA + GPR was chosen as the best model, with the following best values: (T = 3.38E + 02, P = 4.0E + 02, Solubility = 0.001241).

Data availability

All data are available within the published paper.

Received: 23 May 2022; Accepted: 25 July 2022
Published online: 30 July 2022

References

1. Khanna, I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov. Today* **17**, 1088–1102 (2012).
2. Sarkis, M., Bernardi, A., Shah, N. & Papathanasiou, M. M. Emerging challenges and opportunities in pharmaceutical manufacturing and distribution. *Processes* **9**, 457 (2021).
3. Zhuang, W., Hachem, K., Bokov, D., Ansari, M. J. & Nakhjiri, A. T. Ionic liquids in pharmaceutical industry: A systematic review on applications and future perspectives. *J. Mol. Liquids* **349**, 118145 (2021).
4. Birmingham, B. & Buvanendran, A. 40 - Nonsteroidal Anti-inflammatory Drugs, Acetaminophen, and COX-2 Inhibitors. In *Practical Management of Pain (Fifth Edition)* (eds Benzon, H. T. *et al.*) 553-568.e555 (Mosby, 2014).
5. Dallegri, F., Bertolotto, M. & Ottonello, L. A review of the emerging profile of the anti-inflammatory drug oxaprozin. *Expert Opin. Pharmacother.* **6**, 777–785 (2005).
6. Todd, P. A. & Brogden, R. N. Oxaprozin: A preliminary review of its pharmacodynamic and pharmacokinetic properties, and therapeutic efficacy. *Drugs* **32**, 291–312 (1986).
7. Miller, L. Oxaprozin: A once-daily nonsteroidal anti-inflammatory drug. *Clin. Pharm.* **11**, 591–603 (1992).
8. Hojjati, M., Yamini, Y., Khajeh, M. & Vatanara, A. Solubility of some statin drugs in supercritical carbon dioxide and representing the solute solubility data with several density-based correlations. *J. Supercrit. Fluids* **41**, 187–194 (2007).
9. Foster, N. *et al.* Processing pharmaceutical compounds using dense gas technology. *Ind. Eng. Chem. Res.* **42**, 6476–6493 (2003).
10. Güçlü-Üstündağ, Ö. & Temelli, F. Solubility behavior of ternary systems of lipids, cosolvents and supercritical carbon dioxide and processing aspects. *J. Supercrit. Fluids* **36**, 1–15 (2005).
11. Paul, D. *et al.* Artificial intelligence in drug discovery and development. *Drug Discov. Today* **26**, 80 (2021).
12. Yang, J., Du, Q., Ma, R. & Khan, A. Artificial intelligence simulation of water treatment using a novel bimodal micromesoporous nanocomposite. *J. Mol. Liq.* **340**, 117296 (2021).
13. El Naqa, I. & Murphy, M.J. What is machine learning?. In: Machine Learning in Radiation Oncology. 3–11 (Springer, 2015).
14. Wang, H., Lei, Z., Zhang, X., Zhou, B. & Peng, J. Machine learning basics. *Deep Learn*. 98–164 (2016).
15. Dietterich, T.G. Ensemble methods in machine learning. In: International workshop on multiple classifier systems. 1–15 (Springer, 2000).
16. Zhou, Z.-H. Ensemble methods: Foundations and algorithms, Chapman and Hall/CRC, 2019.
17. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
18. Mathuria, M. Decision tree analysis on j48 algorithm for data mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**, 1114–1119 (2013).
19. Sakar, A. & Mammone, R. J. Growing and pruning neural tree networks. *IEEE Trans. Comput.* **42**, 291–299 (1993).
20. Frau, L., Susto, G. A., Barbariol, T. & Feltresi, E. Uncertainty estimation for machine learning models in multiphase flow applications. *Informatics.* **8**, 58 (2021).
21. Mosavi, A. *et al.* susceptibility mapping of soil water erosion using machine learning models. *Water* **12**, 1995 (2020).
22. Khoshmaram, A. *et al.* Supercritical process for preparation of nanomedicine: Oxaprozin case study. *Chem. Eng. Technol.* **44**, 208–212 (2021).
23. Quinonero-Candela, J. & Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6**, 1939–1959 (2005).
24. Jiang, Y., Jia, J., Li, Y., Kou, Y. & Sun, S. Prediction of gas-liquid two-phase choke flow using Gaussian process regression. *Flow Meas. Instrum.* **81**, 102044 (2021).
25. Quinlan, J. R. Learning decision tree classifiers. *ACM Comput. Surv. (CSUR)* **28**, 71–72 (1996).
26. Xu, M., Watanachaturaporn, P., Varshney, P. K. & Arora, M. K. Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.* **97**, 322–336 (2005).
27. Kushwah, J. S. *et al.* Comparative study of regressor and classifier with decision tree using modern tools. *Mater. Today Proc.* **56**, 3571–3576 (2021).
28. Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. Classification and regression trees. (Routledge, 2017).
29. Segal, M. R. & Bloch, D. A. A comparison of estimated proportional hazards models and regression trees. *Stat. Med.* **8**, 539–550 (1989).
30. Schapire, R.E. The boosting approach to machine learning: An overview. Nonlinear estimation and classification 149–171 (2003).
31. Ying, C., Qi-Guang, M., Jia-Chen, L. & Lin, G. Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica* **39**, 745–758 (2013).
32. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning. Springer series in statistics (Springer, 2001).
33. Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning* Vol. 4. (New York: springer, 2006).
34. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class adaboost, statistics and its. *Interface* **2**, 349–360 (2009).
35. Berk, R. A. An introduction to ensemble methods for data analysis. *Sociol. Methods Res.* **34**, 263–295 (2006).
36. Ouyang, Z., Ravier, P. & Jabloun, M. STL decomposition of time series can benefit forecasting done by statistical methods but not by machine learning ones. *Eng. Proc.* **5**(1), 42 (2021).
37. De Myttenaere, A., Golden, B., Le Grand, B. & Rossi, F. Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48 (2016).
38. Paula, M., Marilaine, C., Nuno, F. J. & Wallace, C. Predicting long-term wind speed in wind farms of northeast brazil: A comparative analysis through machine learning models. *IEEE Lat. Am. Trans.* **18**, 2011–2018 (2020).
39. Botchkarev, A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. Available at SSRN 3177507. (2018).
40. Knez, Z., Skerget, M., Sencar-Bozic, P. & Rizner, A. Solubility of nifedipine and nitrendipine in supercritical CO2. *J. Chem. Eng. Data* **40**, 216–220 (1995).

Acknowledgements

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R145), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This study is part of an award-winning large scale research project financed by Taif University Researcher Supporting Project Number (TURSP-2020/117). The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4290565DSR46). 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 (RGP.2/50/43).

Author contributions

W.K.A.: supervision, original draft writing, modeling, validation. S.M.E: original draft writing, investigation, data analysis. K.A.I.: original draft writing, investigation, data analysis. S.A.: original draft writing, investigation,

data analysis, validation. A.A.: writing, data analysis, software, editing. B.H.: writing, investigation, data analysis. A.D.A.: Writing, investigation, software, editing. A.M.A.: original draft writing, investigation, data analysis. Editing K.V.: Writing, investigation, data analysis. M.A.S.A.: original draft writing, supervision, modeling, validation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.K.A., A.M.A. or M.A.S.A.

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

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