Arabian Journal of Chemistry (2022) **15**, 103502

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Machine learning simulation of pharmaceutical solubility in supercritical carbon dioxide: Prediction and experimental validation for busulfan drug

Arash Sadeghi [a], Chia-Hung Su [b,*], Afrasyab Khan [c], Md Lutfor Rahman [d,*], Mohd Sani Sarjadi [d], Shaheen M. Sarkar [e]

[a] *Research and Development Department, Pars Alcohol Company, Eghlid, Fars, Iran*
[b] *Department of Chemical Engineering, Ming Chi University of Technology, New Taipei City, Taiwan*
[c] *Research Institute of Mechanical Engineering, Department of Vibration Testing and Equipment Condition Monitoring, South Ural State University, Lenin prospect 76, Chelyabinsk 454080, Russian Federation*
[d] *Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia*
[e] *Department of Applied Science, Technological University of the Shannon, Moylish Park, Limerick V94 EC5T, Ireland*

Received 17 July 2021; accepted 10 October 2021
Available online 16 October 2021

KEYWORDS

Artificial intelligence;
Simulation;
Modeling;
Pharmaceutics;
Nanomedicine

Abstract An artificial intelligence-based predictive model was developed using a support vector machine to investigate the solubility data of the drug Busulfan drug in supercritical carbon dioxide. The data for simulations were collected from literature. The model was trained and implemented in order to determine the correlation between the solubility values and the input parameters, namely, temperature and pressure. These parameters were used as the inputs as they are known to have a significant effect on the solubility of Busulfan in supercritical carbon dioxide. In the artificial intelligence model, a polynomial model with kernel function was applied to the data, and the model's findings were compared with measured data for fitting. Good agreement was observed between the model's outputs and the measured data with coefficient of determination greater than 0.99.

1. Introduction

Development of predictive models for pharmaceutical manufacturing is of great importance, and extremely useful for moving towards a Quality-by-Design (QbD) paradigm which is very important for the next generation of pharmaceutical processing (Shirazian, 2017; Ismail, 2019; Shirazian, 2018). In this paradigm, the process of pharmaceutical manufacturing should be thoroughly understood for advanced production and quality assurance. Indeed, the models can help to develop

* Corresponding authors.
E-mail addresses: chsu@mail.mcut.edu.tw (C.-H. Su), lotfor@ums.edu.my (M. Lutfor Rahman).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.arabjc.2021.103502
1878-5352

pharmaceutical processing specifically for solid-dosage oral formulations which constitute the majority of the produced pharmaceuticals (Lou et al., 2021; Stranzinger, 2021). The predictive models can be developed at different levels and for various pharmaceutical unit operations such as granulation, crystallization, milling, coating and reactions (Shirazian, 2017; Walsh et al., 2020; Singh, 2020; Pishnamazi, 2021; Shirazian, 2018; Ismail, 2019; Ismail, 2020). Either mechanistic or non-mechanistic models can be developed for prediction of the performance of pharmaceutical unit operations. The models developed for pharmaceutical processing have been implemented for process development, and great results have been reported. Indeed, robust models need to be developed and implemented for processes.

One of the major challenges in pharmaceutical product development is to improve the solubility and bioavailability of drug substances in aqueous media. This can be achieved by various physio-chemical techniques. One of the techniques for improving drug solubility is nanonization in which the drug particles are manufactured at small scales such that the solubility is increased considerably due to higher surface area and energy of the nanoparticles. The methods of preparation of nanomedicines are diverse but can be divided into two categories of top-down and bottom-up approaches. The bottom-up method has attracted much attention due to the precise control over the product quality. In this method, the drug substance is usually dissolved in a proper solvent, and then the nanosize powder of the drug is formed in a solvent removal process depending on the structure of the drug and operating conditions. However, other methods of nanonization have been reported in the literature (Padrela, 2018).

When developing advanced processes for the production of nano drugs, a sound understanding of the basic process methods is vital, indeed underpinning research is required to understand the formation of nanomedicine in the specific nanonization process used. By understanding the process, one can optimize the production to achieve the best quality and minimize the cost and energy for production of nanomedicine. The models based on mechanics and statistical models can be used and developed for simulation and understanding of nanomedicine production. Basically, before deploying the process, the solubility of drug in the solvent must be determined, as it plays a crucial role in a processing approach based on a bottom-up technique. Recently, some empirical and thermodynamic models have been developed for simulation of drug solubility in supercritical solvent which is used for preparation of nanomedicines.

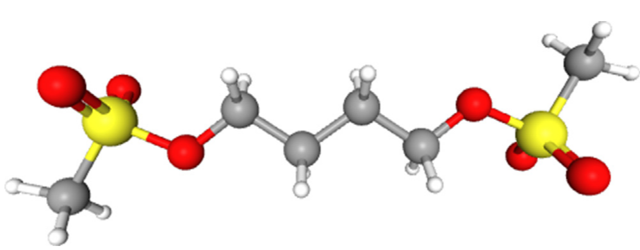

Fig. 1 Chemical structure of the drug used in this study (https://pubchem.ncbi.nlm.nih.gov/compound/Busulfan#section=3D-Conformer. Accessed July 2021).

Table 1 The solubility data of busulfan used in the simulations (Pishnamazi, 2020).

P(bar)	T(K)	Y (mole fraction)
120	3.08E + 02	8.03E-05
120	3.18E + 02	5.17E-05
120	3.28E + 02	4.69E-05
120	3.38E + 02	3.27E-05
160	3.08E + 02	1.26E-04
160	3.18E + 02	1.19E-04
160	3.28E + 02	9.19E-05
160	3.38E + 02	8.91E-05
200	3.08E + 02	1.49E-04
200	3.18E + 02	1.72E-04
200	3.28E + 02	2.10E-04
200	3.38E + 02	2.12E-04
240	3.08E + 02	1.72E-04
240	3.18E + 02	2.10E-04
240	3.28E + 02	2.46E-04
240	3.38E + 02	3.12E-04
280	3.08E + 02	1.97E-04
280	3.18E + 02	2.73E-04
280	3.28E + 02	3.48E-04
280	3.38E + 02	4.40E-04
320	3.08E + 02	2.26E-04
320	3.18E + 02	3.26E-04
320	3.28E + 02	4.27E-04
320	3.38E + 02	5.45E-04
360	3.08E + 02	2.45E-04
360	3.18E + 02	3.46E-04
360	3.28E + 02	4.90E-04
360	3.38E + 02	6.33E-04
400	3.08E + 02	2.74E-04
400	3.18E + 02	3.71E-04
400	3.28E + 02	6.18E-04
400	3.38E + 02	8.65E-04

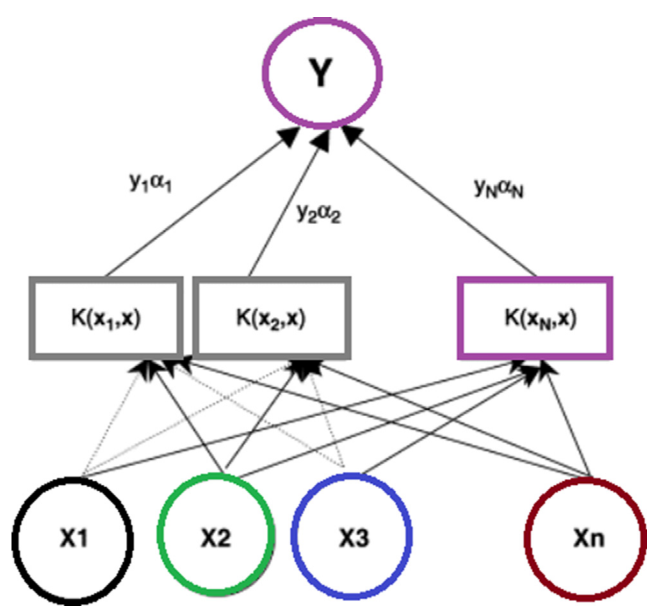

Fig. 2 Schematic structure of SVM model in prediction of busulfan solubility.

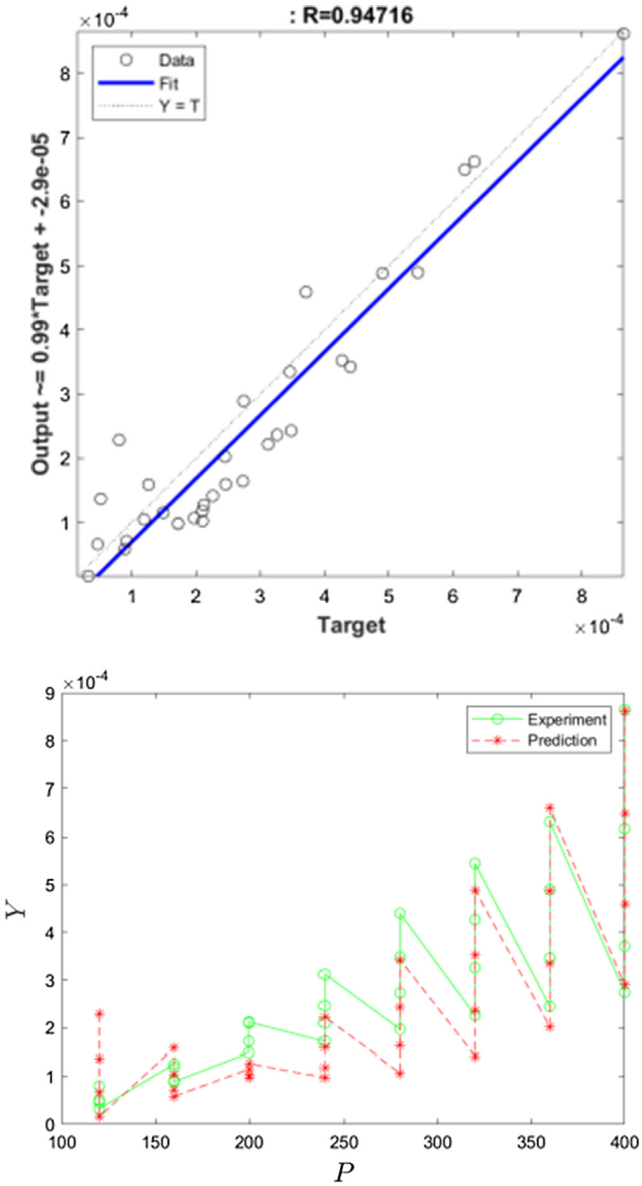

Fig. 3 Prediction of parameter Y based on P and T with SVM model and Kernel Function = polynomial with the order of 2 and box constraint equal to response Scale/0.1.

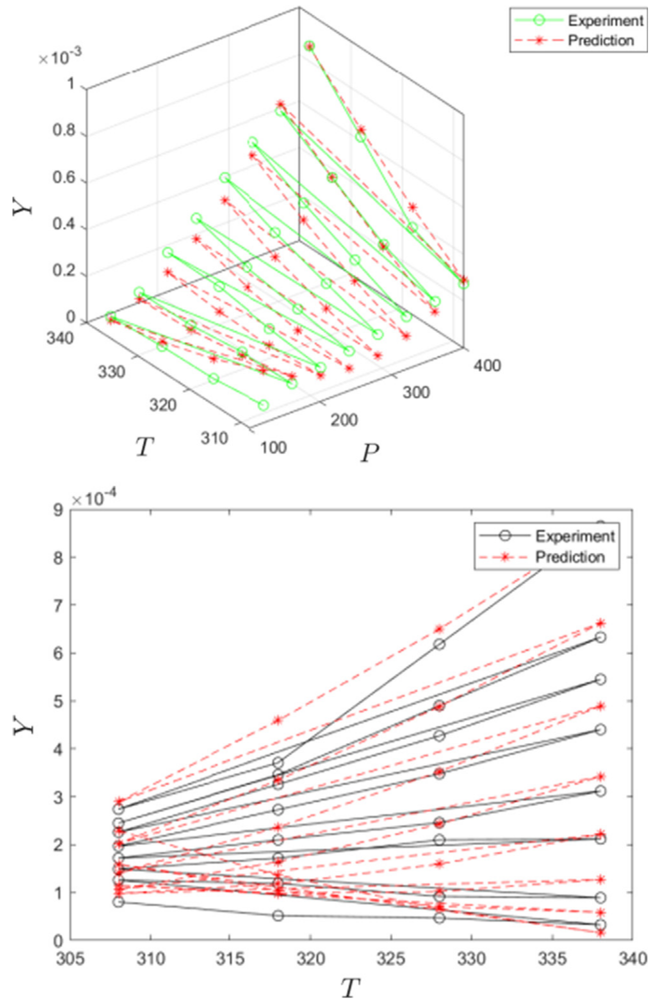

Fig. 3 (*continued*)

Sodeifian et al. (Sodeifian, 2018; Sodeifian, 2020; Sodeifian et al., 2020; Sodeifian et al., 2019) developed a number of thermodynamic-based models for prediction of various drugs solubilities in supercritical CO_2. Zabihi et al. (Pishnamazi, 2021; Pishnamazi, 2021; Pishnamazi, 2021; Zabihi, 2020; Zabihi, 2020; Zabihi, 2021; Pishnamazi, 2020; , xxxx; Zabihi, 2021; Pishnamazi, 2020; Zabihi, 2021; Pishnamazi, 2020) used several semi-empirical correlations for simulation of drug solubility in supercritical solvent. Their results indicated that the employed models are capable of predicting drug solubility in supercritical solvents with high accuracy, and can be used as extrapolative tools for prediction of drug solubility in supercritical carbon dioxide.

The main objective of the current work is to develop a new simulation methodology for prediction of drug solubility at supercritical conditions. The method of simulation is based on support vector machines (SVM) which utilize multi-Class support vector algorithms for prediction and training the data of solubility. The model of drug used in this study is busulfan which is considered a good candidate for the production of nanomedicine in continuous supercritical-based technology. The data are collected from literature and for model's training to build the SVM model. The results of simulation are evaluated in terms of accuracy, and the simulation parameters are tuned to achieve the best prediction of drug solubility. To the best of the authors' knowledge, there is no simulation study on prediction of busulfan solubility in supercritical CO_2 using support vector machine algorithms, which can be considered as the main innovation of the current study.

2. Measurement method

Busulfan with the chemical formula of $C_6H_{14}O_6S_2$, and molecular weight of 246.304 gr/mol was used. The raw drug with purity of 0.98 was treated to remove the impurity, and the drug with the resulting high purity was used in the experiments. We have collected data from the literature, and the detailed description of the drug and measurements can be found elsewhere (Pishnamazi, 2020). The structure of the drug is indicated in Fig. 1.

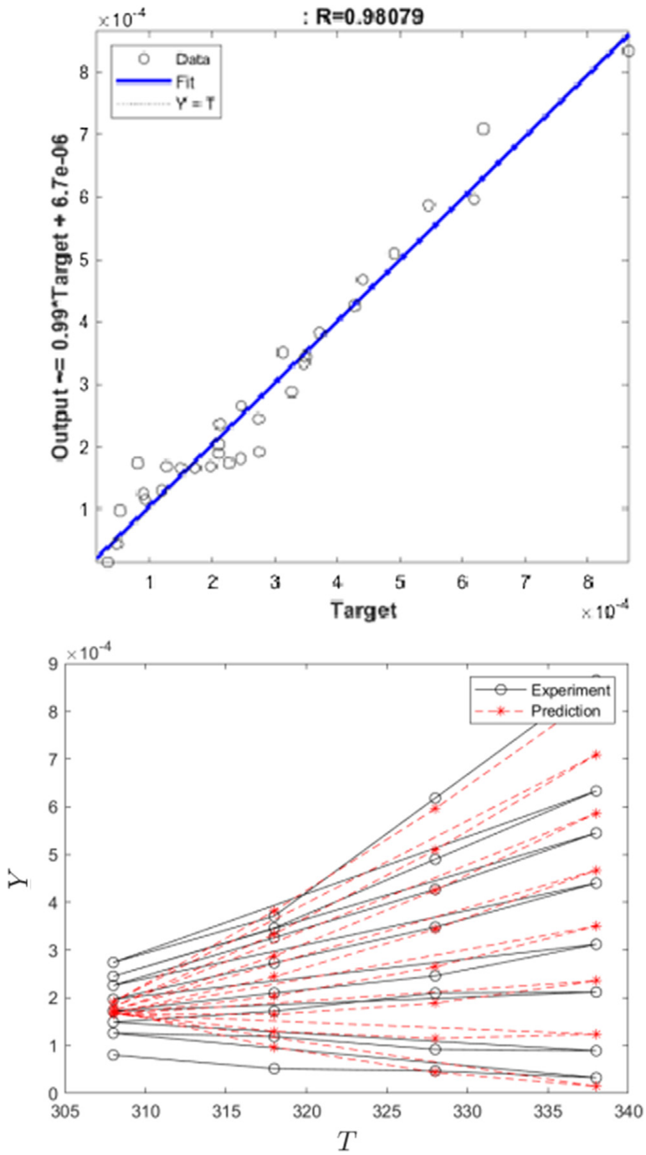

Fig. 4 Prediction of parameter Y based on P and T with SVM model and Kernel Function = polynomial with the order of 2 and box constraint equal to response Scale/2.

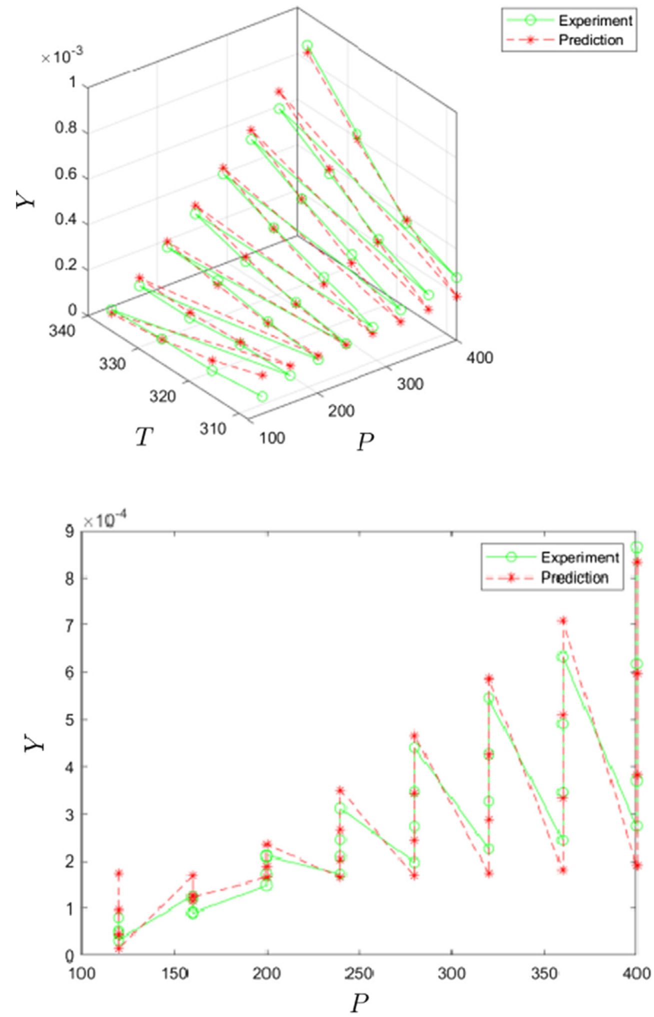

Fig. 4 (*continued*)

The collected solubility data at different pressures and temperatures are listed in Table 1. As seen, 32 values are collected for the drug at 8 pressures and 4 temperatures between 308 and 338 K to evaluate the effect of these two important process parameters on the values of drug solubility. It is worth mentioning that the unit of solubility is mole fraction (Y) in the calculations, between 0 and 1.

3. Method of calculations

For modelling and simulations of the solubility data, a method of support vector machines (SVM) was employed due to the complexity of the system. The method of SVM is widely utilized in applications including signal processing, medical applications, natural language processing, and voice and picture recognition. The SVM's main goal is to separate information components into two groups: types and classes. The algorithm uses a hyperplane to separate linearly separable data sets, however it will implement a soft margin boost when it comes to identifying practical solutions.

SVMs' accuracy for classification and estimation is quite high, and their results on classification and regression tasks are remarkably. A multi-class SVM is made up of several binary classifiers. Kernels help to open up nonlinear issues, giving them more flexibility and capability to deal with different situations. Using training data, we just need support vectors to create a decision surface. The resulting model is fully capable of code generation after the training is complete.

The binary classifier that the SVM generates is termed the optimum splitting hyperplane and is produced via a projection that develops highly nonlinear input variables into the high-dimensional feature space. SVM utilizes non-linear class boundaries in constructing a linear decision model. An SVM trained on linearly split data provides a hyperplane that accurately separates the data, but the separation is accomplished as far away from the learning options as possible. In this study the polynomial machine with kernel function can be defined as:

$$k(x, x_i) = (x.x_i + 1)^d \quad (1)$$

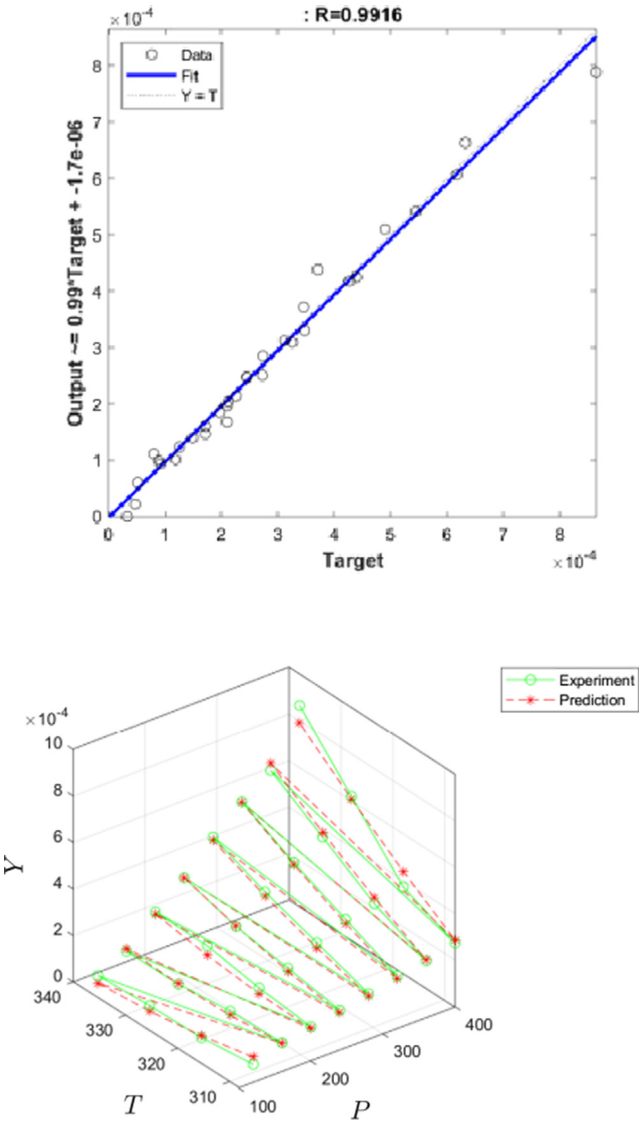

Fig. 5 Prediction of parameter Y based on P and T with SVM model and Kernel Function = polynomial with the order of 2 and box Constraint is equal to response Scale/5.

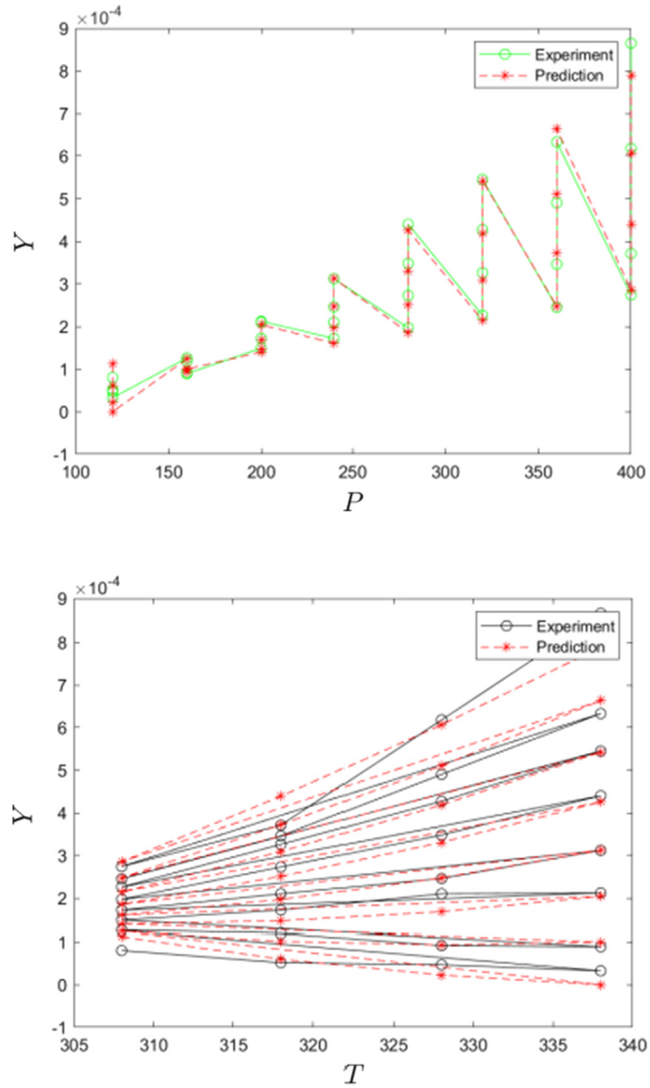

Fig. 5 (*continued*)

where $k(x, x_i)$ & d are kernel function and the degree of the polynomial kernel. The input vector in the method can be defined as $x_i \in R^n$, & $y_i \in (-1, 1)$, i = 1..l and the optimal hyperplane separating the binary decision classes that can be defined in the main algorithm and SVM is:

$$Y = sign(\sum_{i=1} y_i \alpha_i (x.x_i) + b) \quad (2)$$

More details on the structure of SVMs can be found in Fig. 2 (Zhang, 2019).

4. Results and discussion

The Support Vector Machines (SVM) approach is utilized in the present research to simulate the output parameter (Y) that takes into account two distinct inputs, such as P and T. The Y variable is the drug solubility in supercritical CO_2 with the unit of mole fraction (dimensionless). Kernels increase SVM versatility and applicability, allowing them to be deployed in a wide range of situations and settings. The Kernel Function controls polynomial order transformation. As a consequence of this change, the response scale now fits inside the bounds of the box Constraint, improving model performance. The response component of the model is modified to improve prediction capacity and model correctness. Then, the optimal model is chosen, in this instance based on the optimum response scale.

Fig. 3 shows the prediction of Y as a function of P and T for Kernel Function = polynomial of order 2 and box Constraint = response Scale/0.1. The findings indicate that the R value for the model's output and target values is about 0.94, indicating a good degree of prediction capacity. The SVM findings likewise follow the trend of experimental values for the P and T parameters. However, , this model over-predicts the Y for lower P values and under-predicts Y for higher P values (150–350). Then, when P greater than 350, the experimental observation and the SVM prediction model agree well.

The model exhibits improved fitting between experimental and predicted values as the percentage of response decreases.

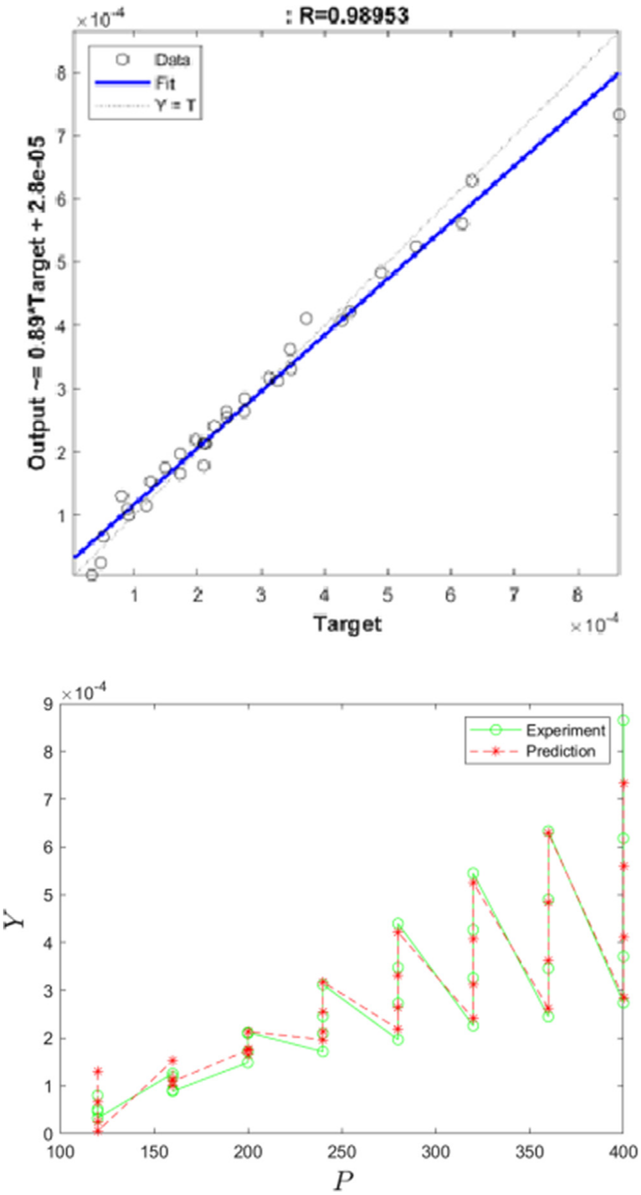

Fig. 6 Prediction of parameter Y based on P and T with SVM model and Kernel Function = polynomial with the order of 2 and box constraint equal to response Scale/10.

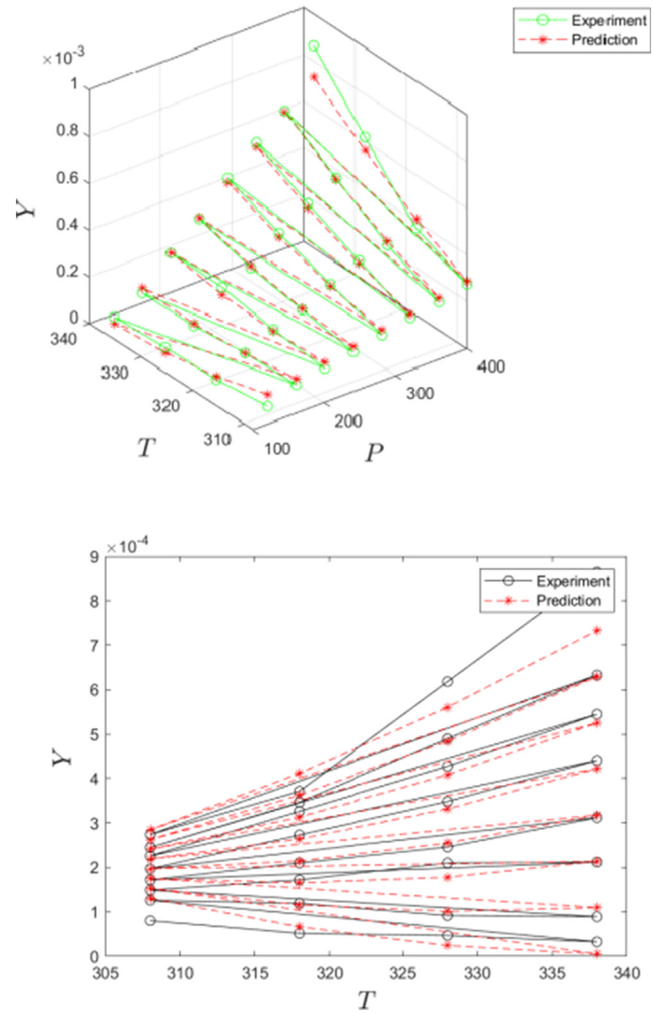

Fig. 6 (*continued*)

In this instance, the R value rises to 0.98, and for P values more than 200, the Y values agree well with the experimental results. Furthermore, we can still observe over-prediction of outcomes for P < 200 (See Fig. 4).

By lowering this ratio, the model's accuracy reaches its maximum prediction capacity. In this instance, the R value is 0.99. Furthermore, the pattern of prediction data sets is completely consistent with experimental results, demonstrating the robustness of these machine learning models (See Fig. 5). More reductions in the box Constraint framework have a detrimental effect on model prediction (See Fig. 6). However, if this decrease is exceeded, the model's accuracy would be completely lost (See Fig. 7).

This SVM-based machine learning model demonstrates that this kind of dataset has excellent modelling potential. However, additional tuning parameters should be included in the training algorithms to improve prediction performance. The new SVM study indicates that the experimental technique may be developed more quickly, with much lower experimental expenses and operational time. In addition, the order of the polynomial varies from 0 to 30 for further tweaking model parameters, and we discovered that the smaller the order of the polynomial, the better the prediction capacity (See Fig. 8).

5. Conclusion

In the current study, a Support Vector Machine (SVM) model is created to simulate the output variable Y, based on the input variables P and T. The model is developed for simulation of a set of solubility data for the busulfan drug in a wide range of pressures and temperatures. The process is simulated for development of supercritical processing in the manufacture of nanomedicine in which supercritical CO_2 is the solvent. The SVM's support for kernel features increases its application and flexibility. The Kernel Function controls all polynomial order transformations. Results show that modification to the Scale Constraint has resulted in improved model performance. Further research is required to determine which tuning may be included into training techniques to improve prediction perfor-

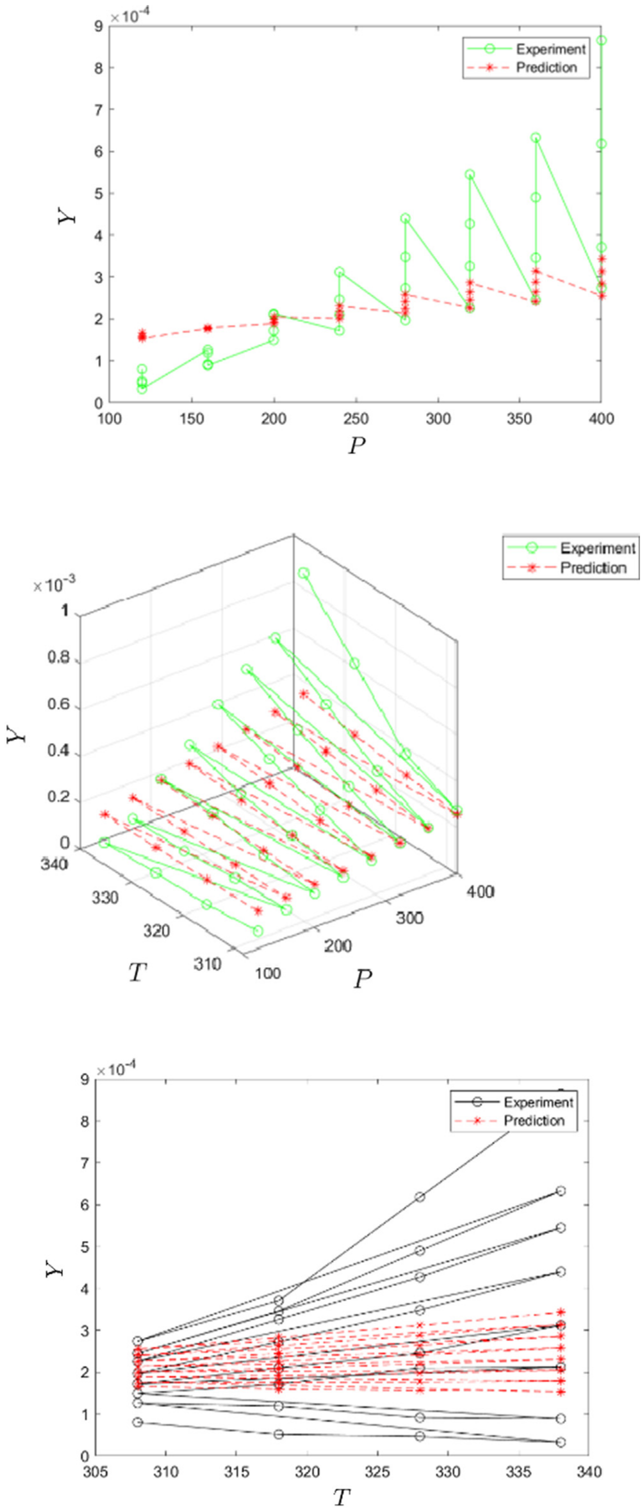

Fig. 7 Prediction of parameter Y based on P and T with SVM model and Kernel Function = polynomial with the order of 2 and box constraint equal to response Scale/20.

mance. This research also demonstrates that machine learning can predict a limited number of datasets. More testing datasets are needed, however, for a better assessment of this approach. Combining SVM with an experimental method could provide

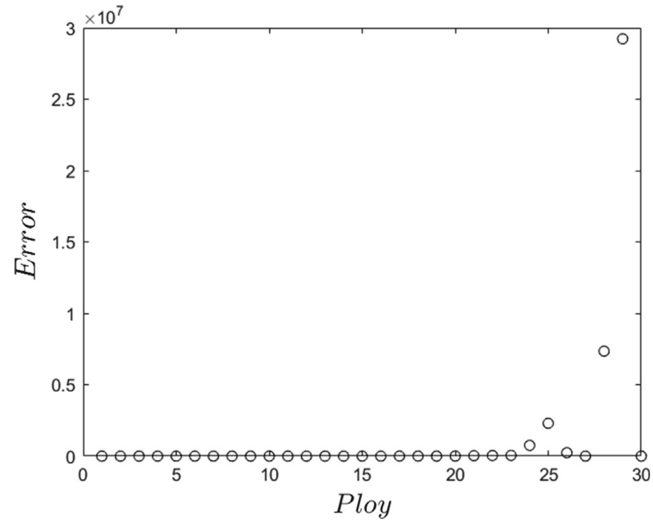

Fig. 8 Error analysis of SVM model and Kernel Function = polynomial with different order and box Constraint is equal to response Scale/0.1.

a continuous domain of findings and results, which saves resources and time by avoiding several experimental runs and the use of costly materials for experimental observation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement:

The authors are thankful to the Russian Government and Research Institute of Mechanical Engineering, Department of Vibration Testing and Equipment Condition Monitoring, South Ural State University, Lenin prospect 76, Chelyabinsk, 454080, Russian Federation for their support to this work

References

Shirazian, S. et al, 2017. Artificial neural network modelling of continuous wet granulation using a twin-screw extruder. Int. J. Pharm. 521 (1–2), 102–109.

Ismail, H.Y. et al, 2019. Developing ANN-Kriging hybrid model based on process parameters for prediction of mean residence time distribution in twin-screw wet granulation. Powder Technol. 343, 568–577.

Shirazian, S. et al, 2018. Regime-separated approach for population balance modelling of continuous wet granulation of pharmaceutical formulations. Powder Technol. 325, 420–428.

Lou, H., Lian, B., Hageman, M.J., 2021. Applications of Machine Learning in Solid Oral Dosage Form Development. J. Pharm. Sci.

Stranzinger, S. et al, 2021. Review of sensing technologies for measuring powder density variations during pharmaceutical solid dosage form manufacturing. TrAC, Trends Anal. Chem. 135, 116147.

Shirazian, S. et al, 2017. Artificial neural network modelling of continuous wet granulation using a twin-screw extruder. Int. J. Pharm. 521 (1), 102–109.

Walsh, J.P., Ghadiri, M., Shirazian, S., 2020. CFD approach for simulation of API release from solid dosage formulations. J. Mol. Liq. 317.

Singh, M. et al, 2020. Characterization of simultaneous evolution of size and composition distributions using generalized aggregation population balance equation. Pharmaceutics 12 (12), 1–17.

Pishnamazi, M. et al, 2021. Chloroquine (antimalaria medication with anti SARS-CoV activity) solubility in supercritical carbon dioxide. J. Mol. Liq. 322, 114539.

Shirazian, S. et al, 2018. Continuous twin screw wet granulation: The combined effect of process parameters on residence time, particle size, and granule morphology. J. Drug Delivery Sci. Technol. 48, 319–327.

Ismail, H.Y. et al, 2019. Developing ANN-Kriging hybrid model based on process parameters for prediction of mean residence time distribution in twin-screw updates wet granulation. Powder Technol. 343, 568–577.

Ismail, H.Y. et al, 2020. Development of high-performance hybrid ANN-finite volume scheme (ANN-FVS) for simulation of pharmaceutical continuous granulation. Chem. Eng. Res. Des. 163, 320–326.

Padrela, L. et al, 2018. Supercritical carbon dioxide-based technologies for the production of drug nanoparticles/nanocrystals – A comprehensive review. Adv. Drug Deliv. Rev. 131, 22–78.

Sodeifian, G. et al, 2018. A comprehensive comparison among four different approaches for predicting the solubility of pharmaceutical solid compounds in supercritical carbon dioxide. Korean J. Chem. Eng. 35 (10), 2097–2116.

Sodeifian, G. et al, 2020. Experimental and thermodynamic analyses of supercritical CO2-Solubility of minoxidil as an antihypertensive drug. Fluid Phase Equilib. 522, 112745.

Sodeifian, G., Sajadian, S.A., Derakhsheshpour, R., 2020. Experimental measurement and thermodynamic modeling of Lansoprazole solubility in supercritical carbon dioxide: Application of SAFT-VR EoS. Fluid Phase Equilib. 507.

Sodeifian, G., Detakhsheshpour, R., Sajadian, S.A., 2019. Experimental study and thermodynamic modeling of Esomeprazole (proton-pump inhibitor drug for stomach acid reduction) solubility in supercritical carbon dioxide. J. Supercrit. Fluids 154.

Pishnamazi, M. et al, 2021. Evaluation of Supercritical Technology for the Preparation of Nanomedicine: Etoricoxib Analysis. Chem. Eng. Technol. 44 (3), 559–564.

Pishnamazi, M. et al, 2021. Experimental and thermodynamic modeling decitabine anti cancer drug solubility in supercritical carbon dioxide. Sci. Rep. 11 (1).

Zabihi, S. et al, 2020. Experimental Solubility Measurements of Fenoprofen in Supercritical Carbon Dioxide. J. Chem. Eng. Data 65 (4), 1425–1434.

Zabihi, S. et al, 2020. Loxoprofen Solubility in Supercritical Carbon Dioxide: Experimental and Modeling Approaches. J. Chem. Eng. Data 65 (9), 4613–4620.

Zabihi, S. et al, 2021. Measuring salsalate solubility in supercritical carbon dioxide: Experimental and thermodynamic modelling. J. Chem. Thermodyn. 152, 106271.

Pishnamazi, M. et al, 2020. Measuring solubility of a chemotherapy-anti cancer drug (busulfan) in supercritical carbon dioxide. J. Mol. Liq. 317, 113954.

Khoshmaram, A., et al., Supercritical process for preparation of nanomedicine: Oxaprozin case study. Chemical Engineering & Technology. n/a(n/a).

Zabihi, S. et al, 2021. Tenoxicam (Mobiflex) Solubility in Carbon Dioxide under Supercritical Conditions. J. Chem. Eng. Data.

Pishnamazi, M. et al, 2020. Thermodynamic modelling and experimental validation of pharmaceutical solubility in supercritical solvent. J. Mol. Liq. 319, 114120.

Zabihi, S. et al, 2021. Thermodynamic study on solubility of brain tumor drug in supercritical solvent: Temozolomide case study. J. Mol. Liq. 321.

Pishnamazi, M. et al, 2020. Using static method to measure tolmetin solubility at different pressures and temperatures in supercritical carbon dioxide. Sci. Rep. 10 (1).

https://pubchem.ncbi.nlm.nih.gov/compound/Busulfan#section=3D-Conformer. Accessed July 2021.

Zhang, Z. et al, 2019. Support Vector Machine for Regional Ionospheric Delay Modeling. Sensors 19 (13), 2947.