Journal of Molecular Liquids 381 (2023) 121805

Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq

An advanced computational method for studying drug nanonization using green supercritical-based processing for improvement of pharmaceutical bioavailability in aqueous media

Hua Xiao Li [a], Uday Abdul-Reda Hussein [b], Ibrahem Waleed [c], Salah Hassan Zain Al-Abdeen [d], Farag M.A. Altalbawy [e], Zainab Hussein Adhab [f], Ahmed Faisal [g], Mohammad Y. Alshahrani [h], Haider Kamil Zaidan [i], Muath Suliman [j], Xiang Ben Hu [a,*]

[a] Shaanxi Institute of International Trade & Commerce, Xi'an, Shaanxi 712000, China
[b] College of Pharmacy, University of Al-Ameed, Iraq
[c] Medical Technical College, Al-Farahidi University, Iraq
[d] Department of Medical Laboratories Technology, AL-Nisour University College, Baghdad, Iraq
[e] National Institute of Laser Enhanced Sciences (NILES), University of Cairo, Giza 12613, Egypt
[f] Department of Pharmacy, Al-Zahrawi University College, Karbala, Iraq
[g] Department of Pharmacy, Al-Noor University College, Nineveh, Iraq
[h] Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, King Khalid University, Abha, Saudi Arabia
[i] Department of Medical Laboratories Techniques, Al-Mustaqbal University College, Hillah, Babylon, Iraq
[j] Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, King Khalid University, Abha, Saudi Arabia

ARTICLE INFO

Keywords:
Drug solubility
Nanomedicine
Random Forest
Extra Tree
KNN

ABSTRACT

In this study, we implemented and compared various non-mechanistic based models for prediction of drug solubility in supercritical solvent. The data were collected from references and the models were built considering various operational circumstances. Small data sets, like the solubility data used in this study, have always been one of the challenges for modeling in machine learning method. In this study, in order to solve the regression problem related to the solubility of drugs, which includes 32 laboratory data, we implemented and studied models that are naturally compatible with very small data like solubility data of drugs in solvents. These models included Random Forest (RF), KNN and Extra Tree (ET). After obtaining the best settings for each model, their final results were compared in terms of accuracy for predicting drug solubility. The ET model had the best result with a score of 0.9999 on the R^2 criterion. Random forests with 0.978 and KNN with 0.972 also had acceptable regression results. Finally, the trained model was used to display and evaluate the effect of input parameters like pressure and temperature on drug solubility to understand the process.

1. Introduction

Process understanding and predictive models are of great importance for process development in various industries such as pharmaceuticals and food. The models can be implemented and trained at various scales such as molecular level, microscopic, mesoscale, macroscale, and plant scale. The model's application and type depend on the process and usage of model for the process [1–3]. For pharmaceutical area, so far different models at disparate scales have been developed and successfully implemented. For solid oral dosage formulation manufacturing, crystallization is the key step, and the primary models for crystallization step is mass transfer, heat transfer, and population balance model (PBM) [4,5]. These models need to be implemented for the process provided that a numerical scheme has been developed and applied. Different numerical schemes such as finite difference, finite element, and finite volume can be applied for numerical solution of process governing equations.

Beside mechanistic models that have been developed and implemented for pharmaceutical processing, the models based on artificial intelligence can be used for this application. Artificial neural network (ANN) model has been successfully implemented for downstream processing of pharmaceutical processing such as granulation and tablet

* Corresponding author.
E-mail address: xiaohuali2015@163.com (X. Ben Hu).

https://doi.org/10.1016/j.molliq.2023.121805
Received 1 March 2023; Received in revised form 27 March 2023; Accepted 3 April 2023
Available online 7 April 2023

Table 1
Solubility data used in modeling [15].

P (bar)	T (K)			
	308	318	328	338
120	8.26×10^{-5}	4.26×10^{-5}	4.04×10^{-5}	1.64×10^{-5}
160	1.33×10^{-4}	1.13×10^{-4}	7.35×10^{-5}	5.96×10^{-5}
200	1.53×10^{-4}	1.76×10^{-4}	1.95×10^{-4}	2.22×10^{-4}
240	2.11×10^{-4}	2.26×10^{-4}	2.33×10^{-4}	2.59×10^{-4}
280	2.50×10^{-4}	$3.05 \times 10^{--4}$	3.45×10^{-4}	3.87×10^{-4}
320	2.95×10^{-4}	3.78×10^{-4}	4.40×10^{-4}	5.02×10^{-4}
360	3.28×10^{-4}	4.12×10^{-4}	5.21×10^{-4}	6.04×10^{-4}
400	3.74×10^{-4}	4.55×10^{-4}	6.76×10^{-4}	8.92×10^{-4}

release [6,7]. These artificial intelligence-based models have shown much better performance compared to mechanistic models in terms of fitting accuracy, however these models are applicable when a large amount of data from process is available [8,9]. Indeed, these models are versatile and would be viable to be implemented for pharmaceutical processing for process development.

In pharmaceutical processing, production of drug solid particles at submicron size is of great importance for improving drug solubility, and consequently drug efficacy. Production of drugs with high efficacy can improve patient compliance by reducing the drugs side effects. One of the techniques that can be used for production of nanomedicine is supercritical based processing which is also considered as green technology for preparation of nanodrugs [10]. In this new green technique, measuring and correlation of solubility data is the key step for further process development [11,12]. Prediction of drug solubility can reduce the processing costs as well as analytical costs and time. A model with extrapolative nature can be more applicable for this area. AI based models can be used to predict solubility and optimize the process. For development of these predictive of drugs solubility in the solvent, the data of solubility versus temperature and pressure are required [13,14].

The primary aim for this work is to design and implement a comprehensive methodology for prediction of drug solubility in a supercritical solvent in which the drug model is chloroquine. Herein, the size of input data is small and therefore we need to select the necessary models for forecasting accurately and in proportion to these sizes. Therefore, three linear regression models including random forest (RF), k-nearest Neighbors (KNN), and extreme random tree (ET) are candidates to do so. This is because data with smaller dimensions may have a higher risk of over-fitting, and we selected these models to fit the solubility data for the chloroquine drug in supercritical CO_2. In addition, we need to specify the Hyper-parameters of each machine learning model in the best possible way. Therefore, one of the most important steps of this research is to test the data with different configurations and its effects are discussed accordingly. Solubility data are gathered from the literature and used to fit and validate models.

2. Experimental conditions and data

In this study, we used similar experimental data used in [15] to fit and correlate the machine learning models. However, in order to use such data and that the larger the change interval of one of the data is not involved in its greater impact on the final output, the data mentioned in the next section need to be scaled. This helps us build a better model and has no effect on testing and learning. The data are selected for the solubility of chloroquine as the model drug in the temperature between 308 and 338 K, and the pressure between 120 and 400 bar, as listed in Table 1. The detailed procedure on the solubility measurement and operational conditions can be found in [15].

3. Modeling of process

In this research, we have a regression problem with two inputs and one output: Temperature and Pressure and chloroquine solubility (Y) as our only output, as given in Table 1 obtained from [15].

To obtain an accurate model for predicting the amount of output mentioned above, we have used three different models commonly used on small data sets and compared the results of simulation in order to find the best fitting model for the solubility data. These models included: K Nearest Neighbors (KNN), Extremely Randomized Tree (ET), and Random Forest (RF).

3.1. K Nearest Neighbors (KNN) technique

K-nearest Neighbors (KNN) is a technique for supervised classification and regression that finds particular use in situations in which there is minimal previous knowledge regarding the actual distribution of the data [16]. So, this algorithm can be used for small dataset like our dataset with 32 rows for the solubility data as listed in Table 1. K-NN is an instance-based learning or lazy learning technique in which the function is approximated (not calculated accurately) locally and all computation is postponed until final regression or classification. The k-NN method is a basic ML algorithm that can be used for data prediction [17–19].

Consider X_i as an input vector with p features $\left(x_{i1}, \cdots x_{ip}\right)$, between any two samples,x_i and $x_l (l = 1, 2, \cdots, n)$ The Euclidean distance is calculated as following equation shows:

$$d(X_1, X_l) = \sqrt{\left(x_{i1} - x_{l1}\right)^2 + \cdots + \left(x_{ip} - x_{lp}\right)^2} \quad (1)$$

and the corresponding neighborhood to it as:

$$R_i = \{X \in R^p : d(X, X_i) \leq : d(X, X_m), \forall i \neq m\} \quad (2)$$

Here, each R_i is the clusters of elements with output m, and the set of data points that belong to it is X. The estimated value of the new instance x is the mean value of the k nearest training instances for regression tasks.

3.2. Random Forest (RF) and extreme Random tree (ET)

These two methods are similar, and both are based on decision trees. In this section, we describe them and their differences.

Random Forest is an ensemble tree-based (using decision tree as core) method for both classification and regression [20,21].

The Random Forest (RF) can be used to prevent overfitting in the decision tree. Each tree is trained by drawing a random subset of data from the full training set, and then constructing a decision tree in which each node makes a split based on a feature drawn at random from the entire feature set. Random forest training is very quick, even for large data sets with numerous attributes and tree instances, because each tree is trained separately from the others [22]. The generalization error is accurately approximated by the Random Forest technique, which prevents overfitting [23].

The extreme random tree method was proposed by researchers in [24]. The extreme random tree built a series of "free-growing" regression tree sets using the traditional top-down method. Similar to the RF method, the ET method is also composed of multiple decision trees as core learner. The difference between ET and the random forest method is that the extreme random tree method gets the branching value completely at random to perform the regression tree branching, which is different from the random forest method. Also, each regression tree in the extreme random tree method uses all the training samples.

3.3. Accuracy criteria of models

We utilized three distinct criteria in order to make comparisons, determine which model was superior, and improve the accuracy of the final product. The computed value of the coefficient of determination based on the test data and the training data. The training phase makes

Table 2
List of the accuracy of different configs of KNN model.

K = Number of neighbors	weight function	RMSE	MSE	MAE
5	distance	2.35E−05	5.52E−10	1.99E−05
5	uniform	2.40E−05	5.76E−10	2.24E−05
4	distance	2.59E−05	6.71E−10	2.16E−05
7	distance	2.79E−05	7.78E−10	2.26E−05
6	distance	3.02E−05	9.12E−10	2.56E−05
4	uniform	3.10E−05	9.61E−10	2.63E−05
6	uniform	3.27E−05	1.07E−09	2.83E−05
7	uniform	3.13E−05	9.77E−10	2.45E−05
8	distance	3.10E−05	9.63E−10	2.41E−05
2	distance	3.94E−05	1.55E−09	3.03E−05

use of the remaining two thirds of the data after the test data has been taken up one third of the space in the total data set used for testing. The R^2 score is calculated using Equation (3).

$$R^2 = 1 - \frac{u}{v} \tag{3}$$

where,

$$u = \sum_i (Q_i - y_i)^2 \tag{4}$$

$$v = \sum_i (\bar{y} - y_i)^2 \tag{5}$$

The k-fold cross validation is the third requirement. K-fold is employed to ensure our final method has no overfitting issues.

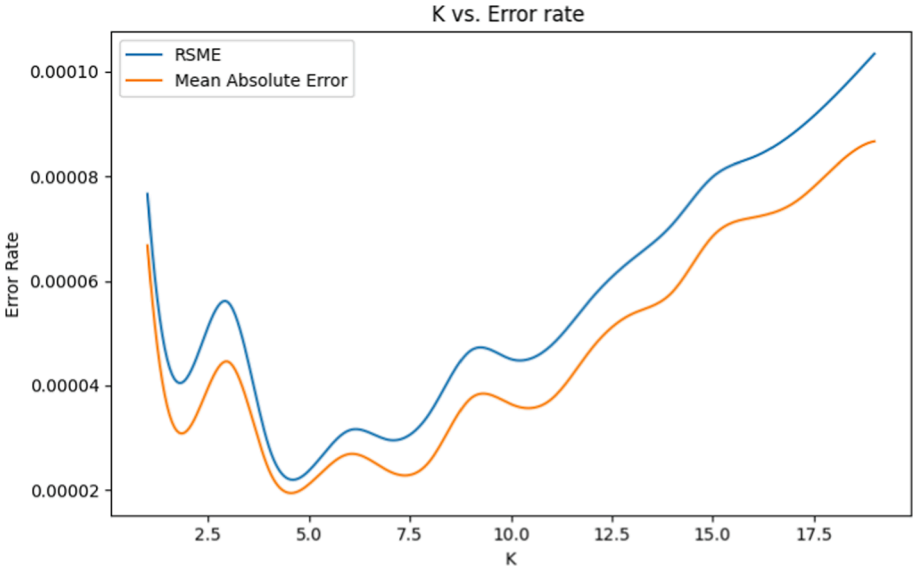

Fig. 1. RMSE and MAE for KNN model.

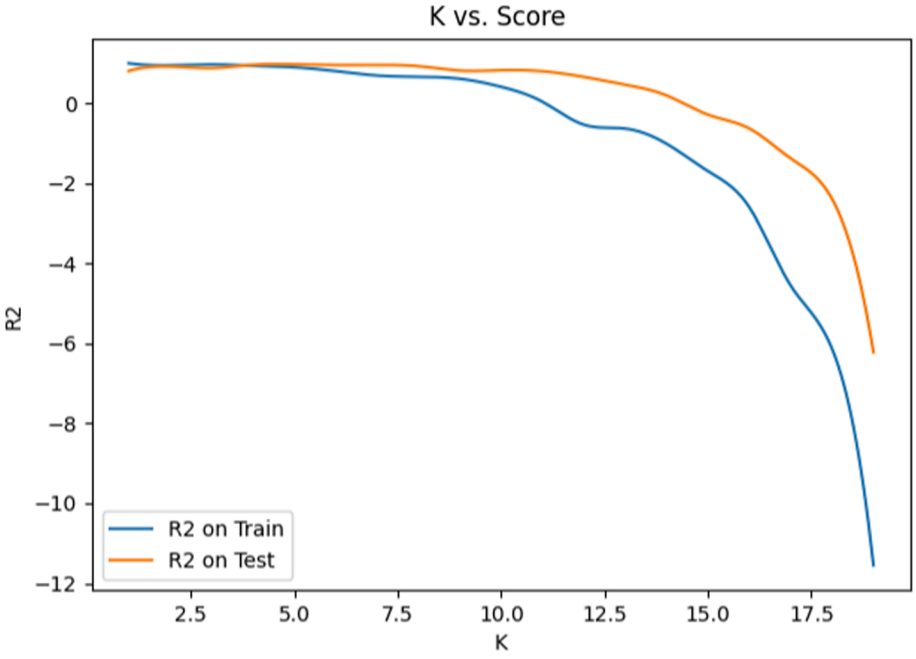

Fig. 2. Evaluating R^2 Score on KNN model.

Table 3
Sample results of RF.

Number of trees	Max Depth	R^2 on Train	RMSE	MSE	MAE	Criterion
7	17	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	11	0.98258	5.99E−05	3.58E−09	3.96E−−05	mae
7	15	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	7	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	9	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	19	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	13	0.98258	5.99E−05	3.58E−09	3.96E−05	mae
7	5	0.98173	6.02E−05	3.63E−09	3.99E−05	mae
5	7	0.97967	4.08E−05	1.66E−09	2.87E−05	mse
5	9	0.97967	4.08E−05	1.66E−09	2.87E−05	mse

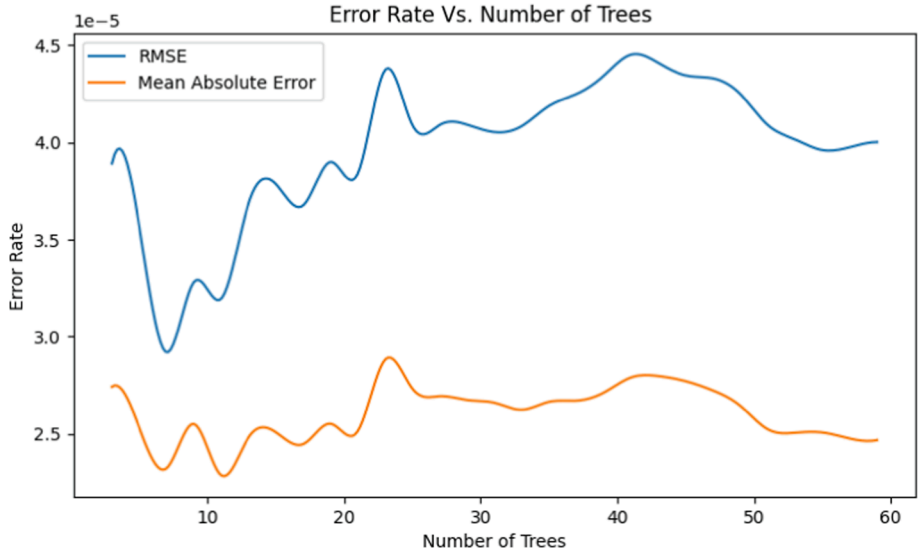

Fig. 3. Variations of accuracy of RF with number of trees changes.

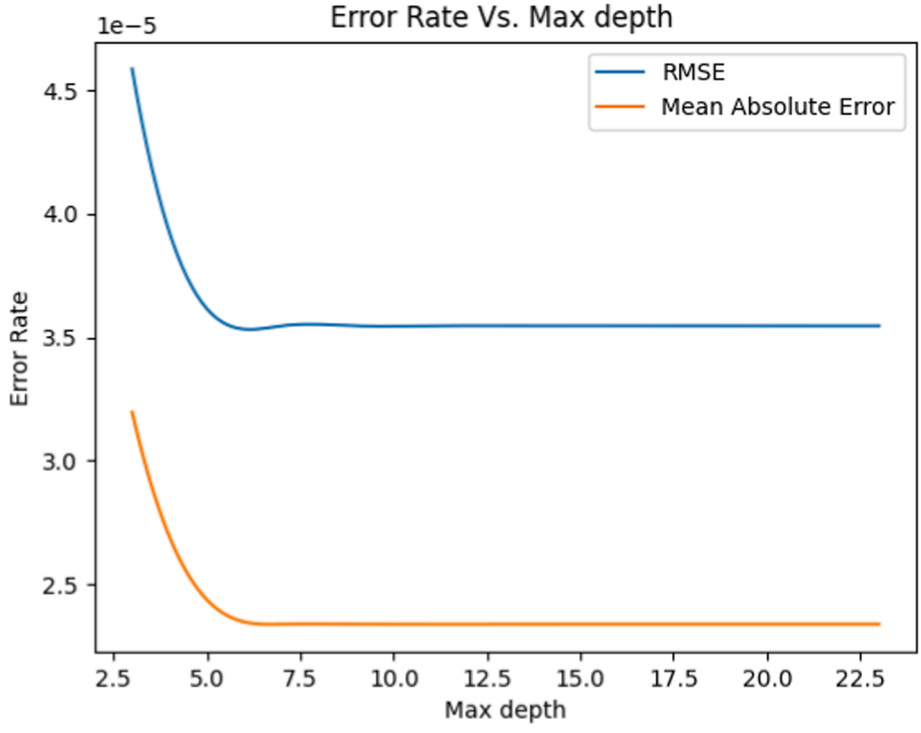

Fig. 4. Variations of accuracy of RF with max depth changes.

Table 4
The statistical analysis results on Extra Tree model.

Number of trees	Max Depth	R^2 on Train	R^2 on Test	RMSE	MSE	MAE
25	7	0.99996	0.98503	1.98E−05	3.91E−10	1.68E−05
25	8	0.99999	0.98455	1.99E−05	3.94E−10	1.64E−05
27	8	1	0.98453	1.98E−05	3.93E−10	1.65E−05
27	7	0.99996	0.98381	2.05E−05	4.21E−10	1.71E−05
29	8	1	0.98369	2.05E−05	4.20E−10	1.74E−05
35	8	1	0.983	2.09E−05	4.38E−10	1.75E−05
35	18	1	0.98291	2.10E−05	4.40E−10	1.78E−05
35	11	1	0.98291	2.10E−05	4.40E−10	1.78E−05
35	19	1	0.98291	2.10E−05	4.40E−10	1.78E−05
35	9	1	0.98291	2.10E−05	4.40E−10	1.78E−05
35	10	1	0.98291	2.10E−05	4.40E−10	1.78E−05

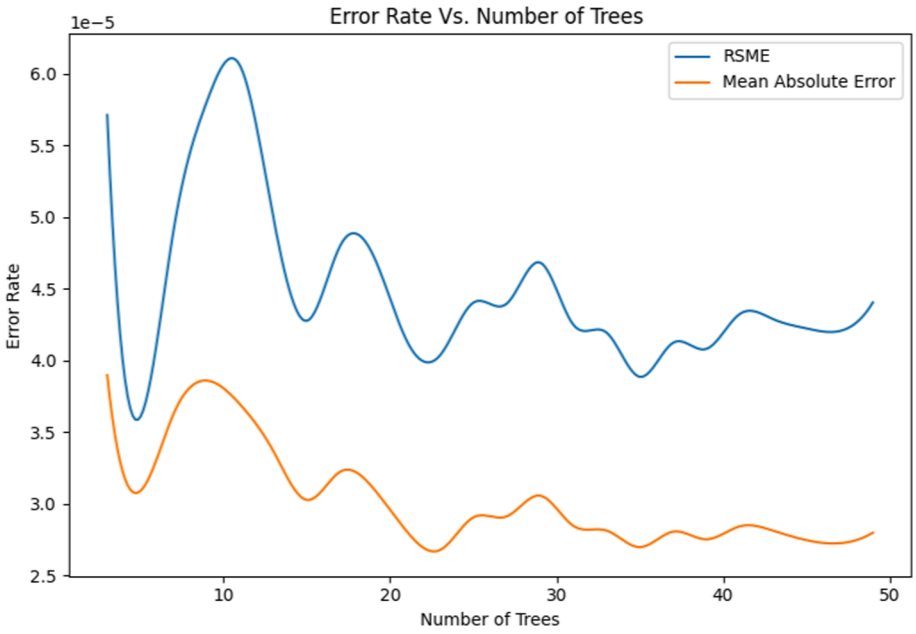

Fig. 5. Effect of No. of trees on fitting error.

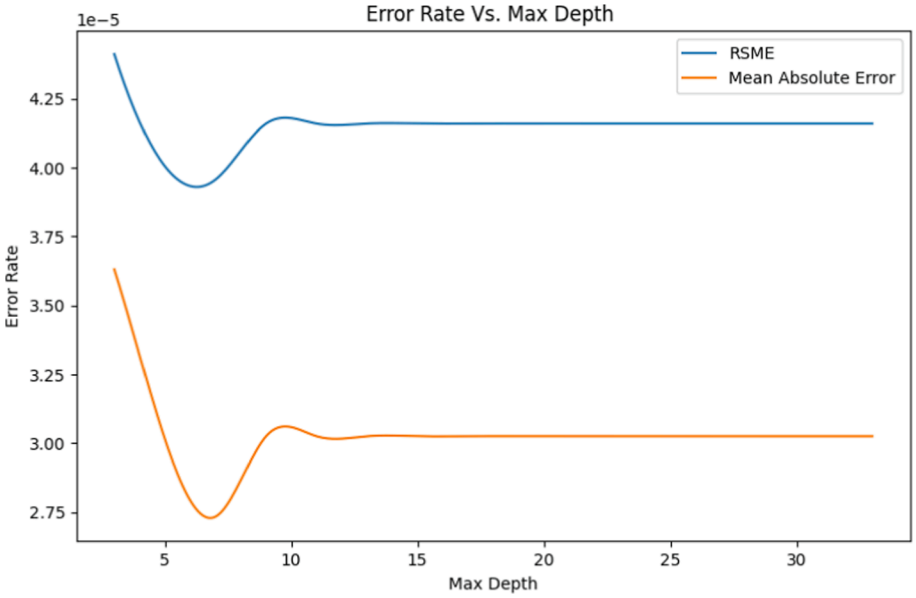

Fig. 6. Effect of Max Depth on fitting error.

Table 5
Best Hyper Parameters for ET model.

Number of trees	Max Depth
25	7

Table 6
Performance of Final Models.

Model	MSE	RMSE	MAE	Train R^2
KNN	5.7588E−10	2.3998E−05	2.24460E−05	0.9728
Extra Tree	4.8572E−10	2.2039E−05	1.92493E−05	0.99997
Random Forest	4.9552E−10	2.2260E−05	1.66986E−05	0.97801

3.4. Choosing the best Hyper-parameters

Now, we need to find the best parameters for models to compare the results. For this aim, different values were tested with our data. For KNN we tried optimizing the K and weight function used in prediction. Table 2 shows an overview of the parameters of KNN.

As it is clear from Figs. 1 and 2, according to all 4 criteria examined, the value of K = 5 is the optimal value for this model. Some of results for Random Forest are listed in Table 3.

Also, in Fig. 3 the impact of changing the quantity of decision trees in Random Forest is shown. With both Fig. 3 and the table, we can find the best number of trees equal to 7.

According to Fig. 4, increasing max depth decreases error rate up to 7. But for more values, there is no effect on the error rate. So, we can

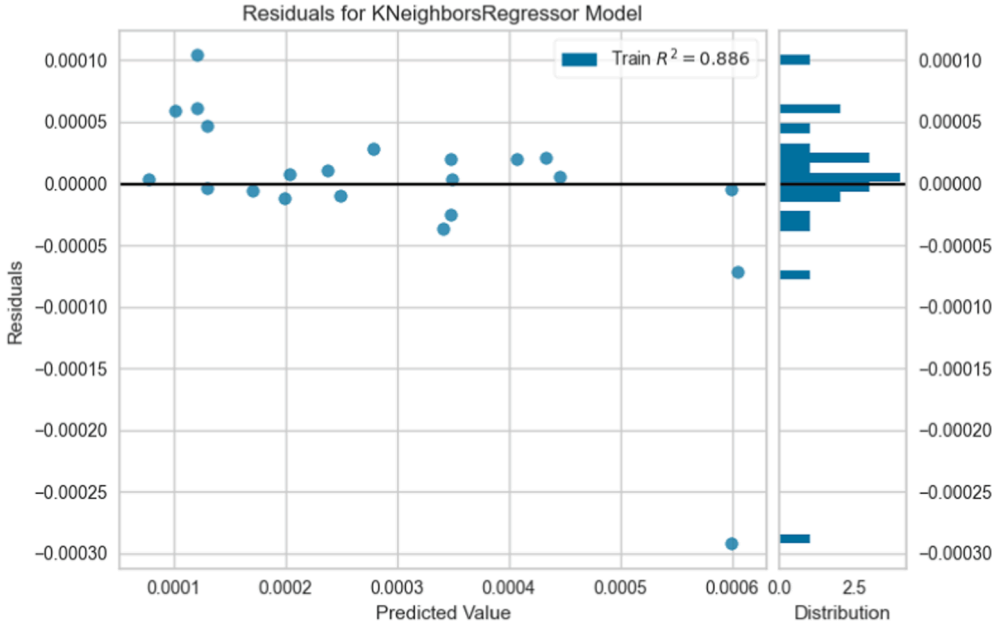

Fig. 7. Residuals with KNN.

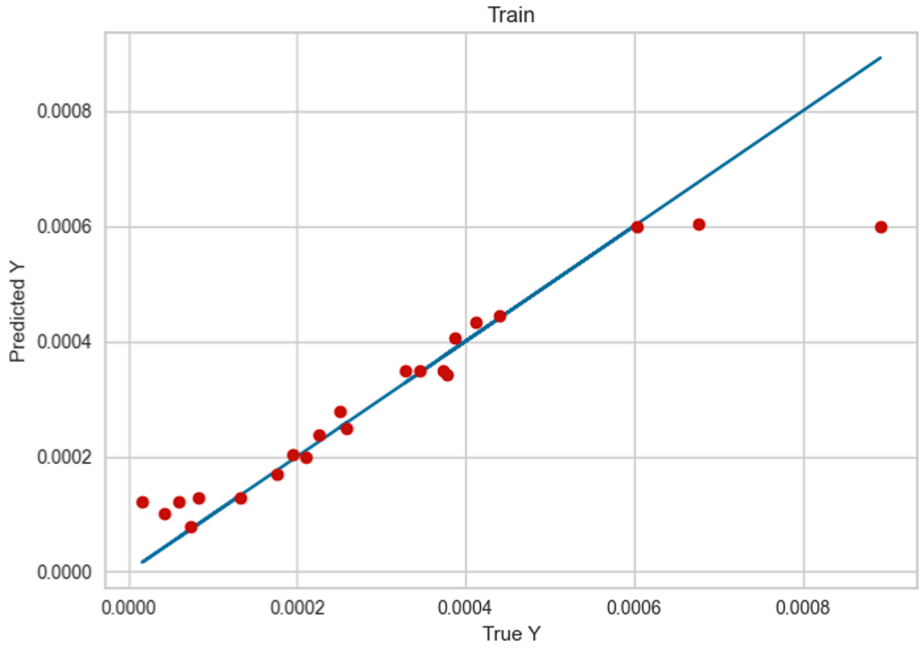

Fig. 8. Comparing Train prediction with true output (KNN Model).

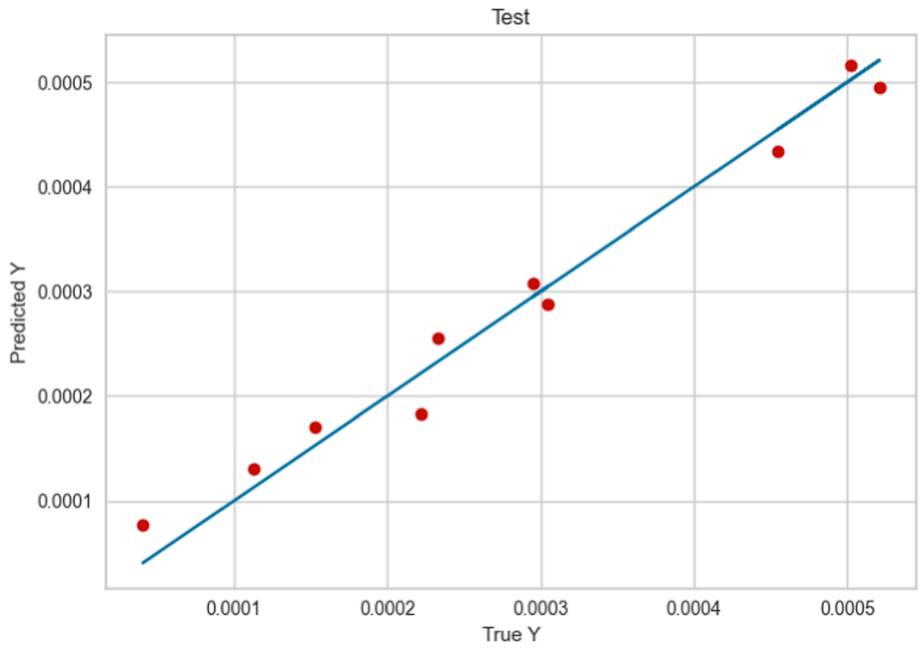

Fig. 9. Comparing Test prediction with true output (KNN Model).

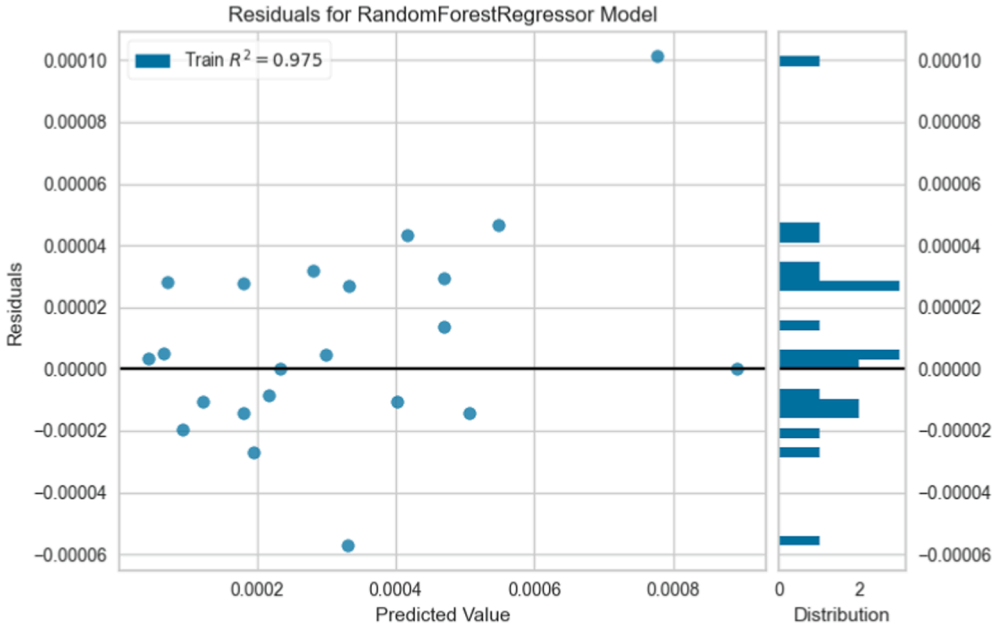

Fig. 10. Residuals with RF.

choose the number of trees = 7 and max depth = 7 for the optimal random forest.

For the ET model, more than 800 different configurations were tested. As we can see in Table 4, the R^2 score in some cases are equal to 1 and this shows that the model operates very accurately in the learning phase (see Figs. 5 and 6). This accurate model hyper parameters are shown in Table 5.

4. Results and discussions

According to last section, models with these hyperparameters are selected to solve our regression problem:

- KNN (Number of neighbors = 5)
- RF (Criterion = mae, N_estimators = 7, Max Depth = 7)
- ET (Criterion = mae, N_estimators = 25, Max Depth = 7)

Table 6 presents the findings obtained from the final models. According to this table we can now analyze these models in advance to evaluate their performance in predicting the drug solubility values.

4.1. KNN results

Final results for KNN with k = 5, it has 0.9728 score in R^2 measurement for fitting the solubility data. This fact shows that this model has a relatively good accuracy considering the size of the data set. The same can be deduced from Fig. 7. However, according to Figs. 8 and 9, in some cases the predicted result is significantly different from the value

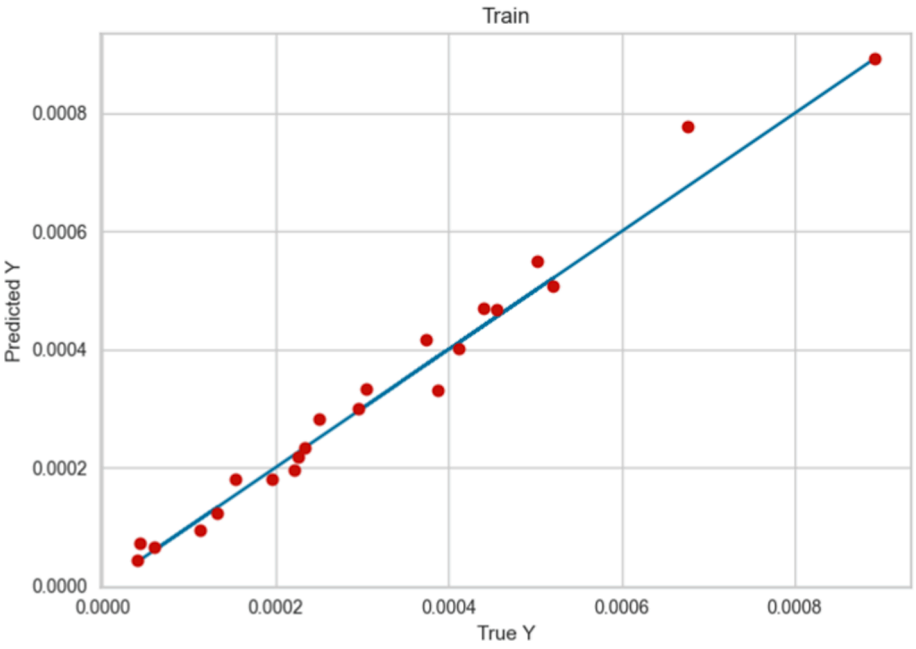

Fig. 11. Comparing Train prediction with true output (RF Model).

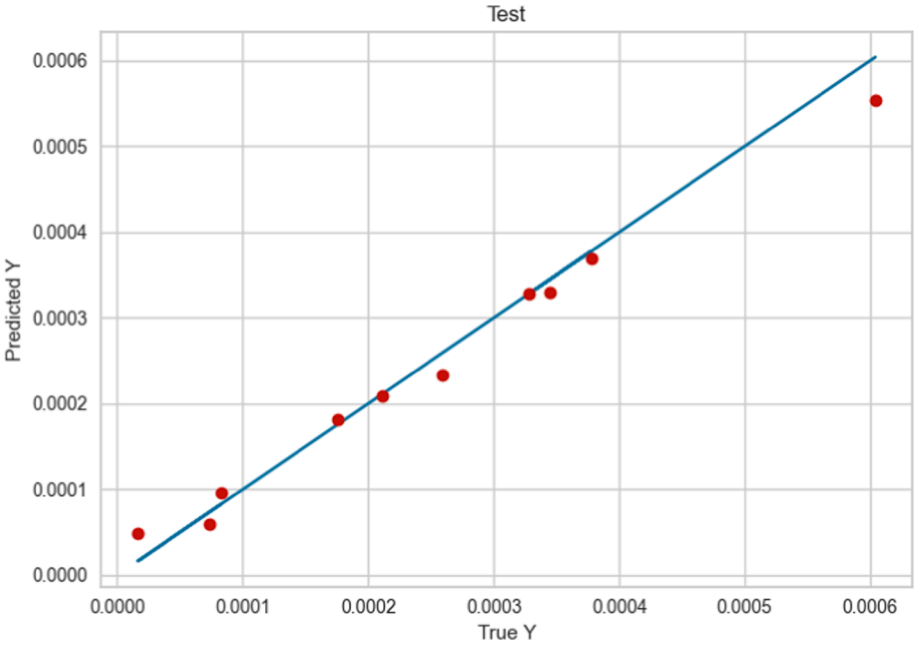

Fig. 12. Comparing Test prediction with true output (RF Model).

observed in the experimental data.

4.2. RF results

Same for Random Forest R^2 score shows good accuracy, but comparing Figs. 10, 11, and 12 with the former subsection, we can see that the RF model is less suspected of over-fitting.

4.3. ET results

As we can see from Table 6, ET can obtain a model that goes through all the examples in the learning phase. This fact is quite clear in Figs. 13 and 14. In addition, according to Fig. 15, we can be sure of the robustness of the model compared to outgoing input data.

Therefore, the ET model with the parameters mentioned at the beginning of Section 4 can be considered the best model available for the problem raised in this research for correlating drug solubility data. Therefore, the predicted solubility values are plotted versus temperature and pressure which are shown in Fig. 16. Pressure, more than temperature, is seen to significantly affect chloroquine solubility, which could be attributed to the compressible behavior of the solvent which is at supercritical state in this process for measuring the solubility.

5. Conclusion

In this investigation, we looked at the issue of solubility using three

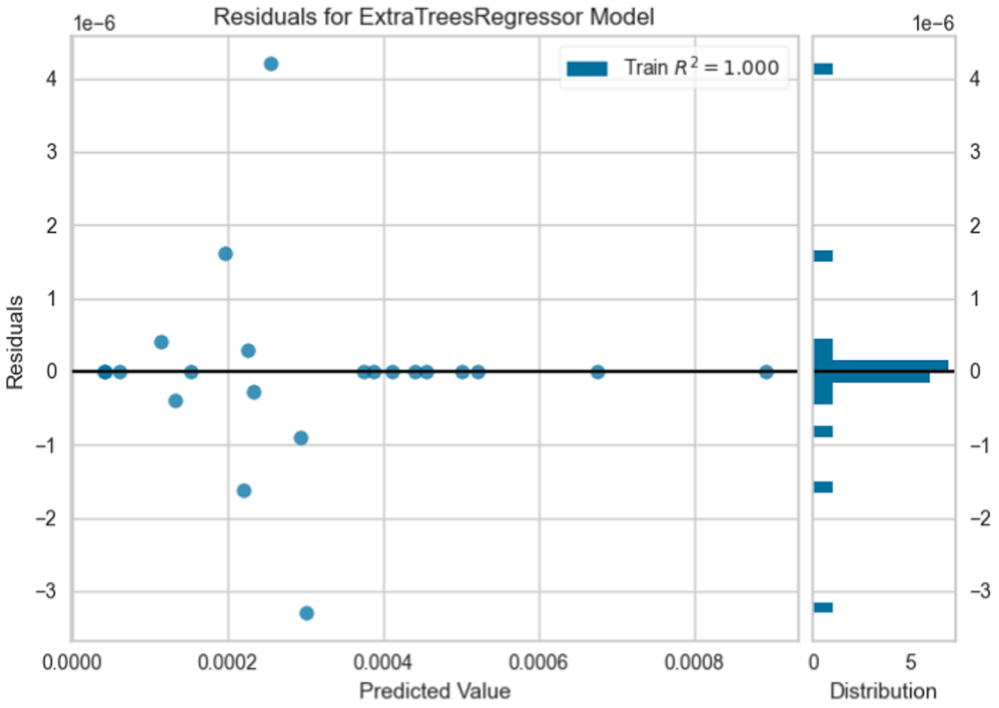

Fig. 13. Residuals with ET.

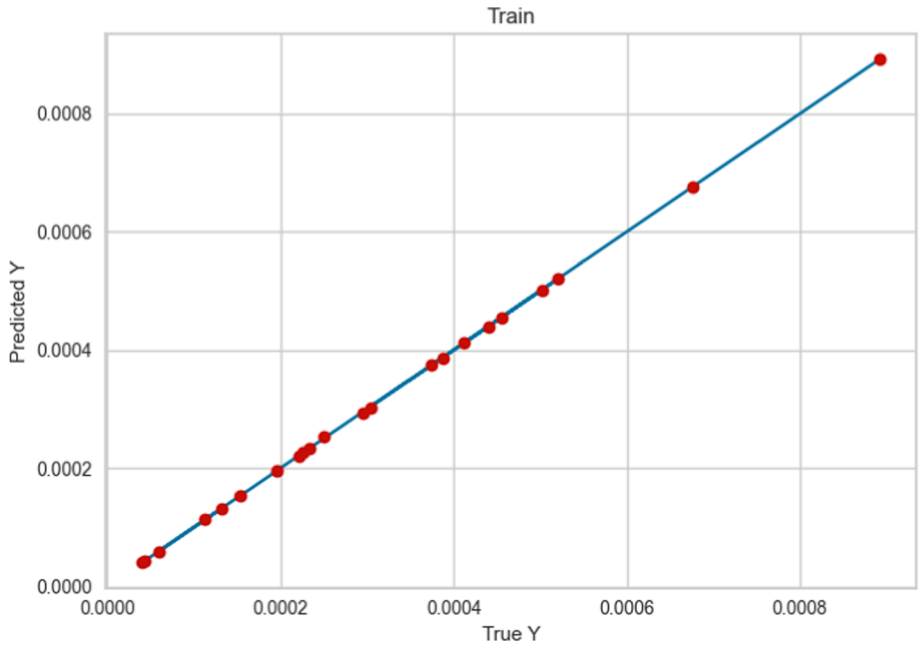

Fig. 14. Comparing Train prediction with true output (ET Model).

different approaches to machine learning models that are naturally suitable for a limited data set. Data were gathered from a wide variety of published sources in order to determine the solubility of chloroquine in supercritical carbon dioxide as a solvent. In terms of accuracy and the impact of pressure and temperature on the solubility, the data and models were examined. After optimizing the hyper-parameters of each, we obtained a final model for them. The results of this study, for which more than 1000 different configurations have been tested, showed that with these methods we can increase the score of the learning and testing stage to 0.9999, which is an ideal model for the problem of interpretation. The model of ET indicated the best results in terms of fitting accuracy.

CRediT authorship contribution statement

Hua Xiao Li: Writing – original draft, Conceptualization, Formal analysis. **Uday Abdul-Reda Hussein:** Validation, Resources, Formal analysis. **Ibrahem Waleed:** Writing – original draft, Investigation, Data curation. **Salah Hassan Zain Al-Abdeen:** Formal analysis, Investigation, Software, Writing – original draft. **Farag M.A. Altalbawy:** Writing – original draft, Formal analysis, Data curation. **Zainab Hussein Adhab:** Writing – review & editing, Validation, Resources. **Ahmed Faisal:**

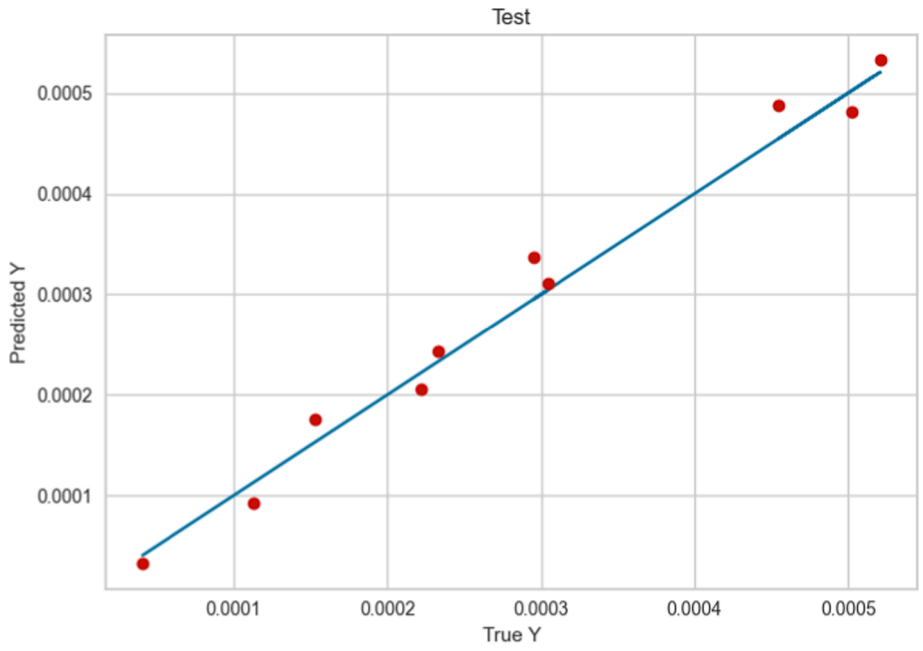

Fig. 15. Compare Test prediction with true output (ET Model).

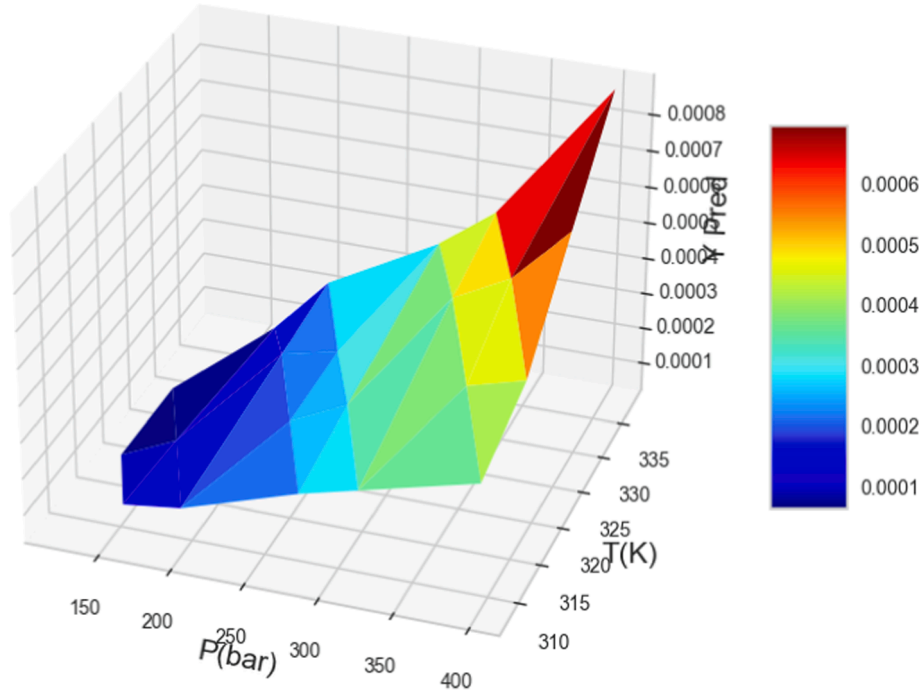

Fig. 16. Surface plot for effect of pressure and temperature on chloroquine solubility with ET (the best model).

Conceptualization, Writing – review & editing, Formal analysis. **Mohammad Y. Alshahrani:** Investigation, Writing – review & editing. **Haider Kamil Zaidan:** Validation, Software, Resources. **Muath Suliman:** Writing – review & editing, Conceptualization, Formal analysis. **Xiang Ben Hu:** Writing – review & editing, Supervision, Funding acquisition.

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.

Data availability

All data are available within the published paper.

Acknowledgements

1. The authors express their gratitude to the Deanship of Scientific Research at King Khalid University for funding this work through the Large Research Group Project under grant number RGP.02/227/43.

2. The research is respectively supported by 4 projects: the project of the General Special Scientific Research of Shaanxi Provincial Education Department (Natural Science Project) in 2022 (22JK0276)

3. a project of the Key Research and Development for Shaanxi Provincial Science and Technology Department in 2022 (2022SF-448).

4. a project of the scientific research project of Shaanxi Institute of International Trade & Commerce (SMXY202219).

References

[1] M. Mazzotti, T. Vetter, and D.R. Ochsenbein, *Crystallization Process Modeling.* Polymorphism in the Pharmaceutical Industry: Solid Form and Drug Development, 2018: p. 285-304.
[2] B.W. Cue, J. Zhang, Green process chemistry in the pharmaceutical industry, Green Chem. Lett. Rev. 2 (4) (2009) 193–211.
[3] M. Trampuž, D. Teslić, B. Likozar, Process analytical technology-based (PAT) model simulations of a combined cooling, seeded and antisolvent crystallization of an active pharmaceutical ingredient (API), Powder Technol. 366 (2020) 873–890.
[4] S.K. Jha, S. Karthika, T.K. Radhakrishnan, Modelling and control of crystallization process, Resour.-Effic. Technol. 3 (1) (2017) 94–100.
[5] Z.K. Nagy, G. Fevotte, H. Kramer, L.L. Simon, Recent advances in the monitoring, modelling and control of crystallization systems, Chem. Eng. Res. Des. 91 (10) (2013) 1903–1922.
[6] D. Barrasso, A. Tamrakar, R. Ramachandran, A reduced order PBM–ANN model of a multi-scale PBM–DEM description of a wet granulation process, Chem. Eng. Sci. 119 (2014) 319–329.
[7] D. Barrasso, A. Tamrakar, R. Ramachandran, Model Order Reduction of a Multi-scale PBM-DEM Description of a Wet Granulation Process via ANN, Procedia Eng. 102 (2015) 1295–1304.
[8] A. Schuhmacher, et al., Big Techs and startups in pharmaceutical R&D – A 2020 perspective on artificial intelligence, Drug Discov. Today 26 (10) (2021) 2226–2231.
[9] X. Tian, H. Tang, L. Cheng, Z. Liao, Y. Li, J. He, P. Ren, M. You, Z. Pang, Evaluation System Framework of Artificial Intelligence Applications in Medical Diagnosis and Treatment, Procedia Comput. Sci. 214 (2022) 495–502.
[10] Y. Cao, et al., Neural simulation and experimental investigation of Chloroquine solubility in supercritical solvent, J. Mol. Liq. 333 (2021), 115942.
[11] Z. Zhao, et al., Multi support vector models to estimate solubility of Busulfan drug in supercritical carbon dioxide, J. Mol. Liq. 350 (2022), 118573.
[12] H. Zhu, 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 (2021), 116731.
[13] W. Liu, et al., Development and validation of machine learning models for prediction of nanomedicine solubility in supercritical solvent for advanced pharmaceutical manufacturing, J. Mol. Liq. 358 (2022), 119208.
[14] S. Xia, Y. Wang, Preparation of solid-dosage nanomedicine via green chemistry route: Advanced computational simulation of nanodrug solubility prediction using machine learning models, J. Mol. Liq. 375 (2023), 121319.
[15] M. Pishnamazi, S. Hosseini, S. Zabihi, F. Borousan, A. Zeinolabedini Hezave, A. Marjani, S. Shirazian, Chloroquine (antimalaria medication with anti SARS-CoV activity) solubility in supercritical carbon dioxide, J. Mol. Liq. 322 (2021) 114539.
[16] N.S. Altman, An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, Am. Stat. 46 (3) (1992) 175–185.
[17] Y. Cai, J.Z. Huang, J. Yin, A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering, Neurocomputing 493 (2022) 191–203.
[18] A.K. Gangwar, A.G. Shaik, k-Nearest neighbour based approach for the protection of distribution network with renewable energy integration, Electr. Pow. Syst. Res. 220 (2023), 109301.
[19] L. Li, X. Chen, C. Song, A robust clustering method with noise identification based on directed K-nearest neighbor graph, Neurocomputing 508 (2022) 19–35.
[20] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[21] G. Biau, Analysis of a Random Forests Model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[22] Y.i. Li, C. Zou, M. Berecibar, E. Nanini-Maury, J.-W. Chan, P. van den Bossche, J. Van Mierlo, N. Omar, Random forest regression for online capacity estimation of lithium-ion batteries, Appl. Energy 232 (2018) 197–210.
[23] N. Khare, S.Y. Sait, Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, Int. J. Pure Appl. Math. 118 (20) (2018) 825–838.
[24] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.