ELSEVIER

Experimental validation and modeling study on the drug solubility in supercritical solvent: Case study on Exemestane drug

Bingfeng Luo [a], Tao Yang [b],*, Sabrean Farhan Jawad [c], Hayder Imad Jabar [d], Hasan Khalid Dabis [e], Mohaned Adil [f], Anfal Nabeel Mustafa [g], Salema K. Hadrawi [h], Ibrahim Mourad Mohammed [i], Abdullah Alshetaili [j], Naseer Mehdi Mohammed [k], Umme Hani [l], Amal M. Alsubaiyel [m]

[a] *Department of Pharmacy, The First People's Hospital of Lanzhou, Lanzhou 730050, China*
[b] *Department of Pharmacy, Gansu Provincial Maternity and Child-Care Hospital, Lanzhou 730050, China*
[c] *Department of Pharmacy, Al-Mustaqbal University College, Hilla, Babylon 51001, Iraq*
[d] *Department of Pharmaceutics, College of Pharmacy, University of Al-Ameed, Karbala, Iraq*
[e] *College of Islamic Science, Ahl Al Bayt University, Kerbala, Iraq*
[f] *College of Pharmacy, Al-Farahidi University, Iraq*
[g] *Department of Pharmacy, AlNoor University College, Bartella, Iraq*
[h] *Refrigeration and Air-conditioning Technical Engineering Department, College of Technical Engineering, The Islamic University, Najaf, Iraq*
[i] *AL-Nisour University College, Baghdad, Iraq*
[j] *Department of Pharmaceutics, College of Pharmacy, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia*
[k] *Mazaya University College, Iraq*
[l] *Department of Pharmaceutics, College of Pharmacy, King Khalid University (KKU), Abha 62529, Saudi Arabia*
[m] *Department of Pharmaceutics, College of Pharmacy, Qassim University, Buraidah 52571, Saudi Arabia*

ARTICLE INFO

ABSTRACT

Green processing based on supercritical solvents has attracted much attention recently in different fields such as pharmaceutical industry due to its superior characteristics. Comprehensive modeling was performed in this study to analyze the preparation of nanomedicine using green supercritical processing. Computational analysis was performed in order to estimate the solubility at different pressures and temperatures. The model was developed based on the input parameters and can estimate the only output of the process which is drug solubility in the supercritical solvent. In this work, we examined how temperature and pressure affect EXE (Exemestane) drug solubility using different tree-based ensemble methods. The models used in this analysis are the Random Forest (RF), the Extremely Randomized Tree (ET), and the Gradient Boosting (GB). Model optimization and hyper-parameter tuning are also accomplished with the aid of Golden eagle optimizer (GEOA). The R^2 values for the test phases of ET, GB, and RF were 0.993, 0.985, and 0.978, respectively. The scores are 0.9945, 0.9758, and 0.9904 in train phases. Specifically, the ET model was chosen since it is the most accurate one. Error rates for this model are 2.317 with MSE, 1.522 with RMSE, and 0.2113 with MAPE.

© 2023 Elsevier B.V. All rights reserved.

1. Introduction

Processing of drug particles of solid oral dosage formulations using supercritical methods have been studied recently for production of nanomedicine with enhanced bioavailability. The method has been also known as the green route for production of drug substances due to lack of organic solvents in the system for preparation of the nanomedicine [1–4]. Extensive research has been done on process design as well as measurement and correlation of solubility data for variety of medicine in supercritical carbon

dioxide as the solvent and some methods have been quite successful for the purpose of modeling and correlation the dataset [5–8]. Thermodynamic-based and machine learning (ML) models are among the most widely applied techniques for estimating drug solubility to input variables such as density, pressure, temperature, etc. [9,10].

The methods of machine learning have been successfully tuned and implemented for fitting of medicine solubility to the pressure and temperature, while this method indicated better performance in terms of statistical parameters such as AIC, RMSE, R^2, etc. when compared with other approaches in solubility correlations such as thermodynamic Equation of State models, known as *EoS* models. The methods of ML require measured dataset of solubility to train

its algorithms, and also for the testing the model. Therefore, having obtained the experimental data of solubility, the ML approach would be a great option for correlation of the solubility data so that the time and costs of measurements can be saved accordingly [11].

Classification, grouping, and regression are just a few of the many tasks for which machine learning methods are now widely used. Ensemble models are a set of methods that are rapidly gaining in popularity. In order to improve the generalization and precision of a single learner, scholars have begun to use ensemble methods, which are collections of models whose predictions are combined into a single model. In this subset of ML algorithms, boosting and bagging are two of the most popular methods [12,13]. Boosting provides weights to the samples frequently and retrains the weak estimators to focus more on the underestimated samples [14,15]. Bagging is the process of randomly extracting feature/sample subsets from a dataset in order to train different weak estimators and increase model diversity [16,17]. This study employed three ensemble estimators, all of which are based on DT (Decision Tree). For this study, we chose Random Forests (RF) and Extremely Random Trees (ET) as bagging methods and Gradient Tree Boosting (GB) as boosting methods.

Gradient boosting is primarily concerned with maximizing the cost function by employing weak learners to generate predictions and combining vulnerable learners to minimize the loss function using an additive model. In this model, the loss function is a measure used in the gradient boosting technique to identify the model's coefficient at fitting the underlying data, and decision trees are employed as core models. In order to "correct" the residuals in the forecasts, regression trees are used because they provide absolute values for splits and their output may be merged together. There are new decision trees introduced into the model at each stage [18,19].

Random Forest (RF) constructs a forest in an ad hoc manner. This woodland is filled with Decision Trees (DTs). Furthermore, there is no link between any of the DT in the RF. When a new sample arrives in the anomaly detection field, each DT in the forest predicts which category the sample should belong to, and the results of all DTs are then aggregated to determine which category the sample belongs to. The training feature vectors are initially randomly sampled by the RF (bootstrap samples). When building a DT, the ideal trait to split can be all or any of the features. As a result, the input samples of each DT differ. As a result, each tree produced is distinct. The benefits of randomness avoid overfitting, while using the mean of all DTs for prediction eliminates certain inaccuracies [15,20].

Extra Trees, sometimes known as Extremely Randomized Trees, is an ensemble method based on Decision Tree estimators. When dividing a tree node, the cut point and associated features are produced at random. An Extra Tree is a powerful tool for both classification and regression problems [21,22].

In this work, for the first time, we have employed multiple advanced machine learning models to predict the equilibrium solubility of the drug namely EXE (Exemestane) in supercritical carbon dioxide (Sc-CO_2) at different temperatures and pressures via three methods of Random Forest (RF), Extremely Randomized Tree (ET), and the Gradient Boosting (GB). The outputs of these three models are then compared with measured data for the sake of validation of the results, and the effects of input parameters will be evaluated on the EXE solubility in the solvent.

2. Data of EXE solubility in the solvent

In this study, we are working with a regression task which is the correlation of a drug solubility to the input parameters as listed in Table 1. There are 45 data points in this activity, which are

Table 1
EXE solubility data used for ML modeling [23].

T (K)	P (MPa)	S ($\times 10$ g L^{-1})
308	12.2	0.67
	15.2	1.38
	18.2	1.47
	21.3	2.41
	24.3	2.5
	27.4	3.41
	30.4	4.1
	33.4	4.55
	35.5	5.92
318	12.2	0.56
	15.2	4.01
	18.2	5.13
	21.3	8.23
	24.3	10.39
	27.4	12.33
	30.4	14.78
	33.4	16.47
	35.5	17.78
328	12.2	0.53
	15.2	3.25
	18.2	8.23
	21.3	12.28
	24.3	16.46
	27.4	22.85
	30.4	28.02
	33.4	33.18
	35.5	36.94
338	12.2	0.35
	15.2	3.82
	18.2	8.99
	21.3	16.57
	24.3	24.65
	27.4	35.36
	30.4	45.7
	33.4	59.79
	35.5	68.25
348	12.2	0.34
	15.2	2.5
	18.2	8.42
	21.3	19.2
	24.3	35.36
	27.4	51.34
	30.4	67.31
	33.4	91.74
	35.5	102.67

arranged as follows: There exist two inputs of the model which are numeric (Pressure and Temperature) and one output which is also numeric (Solubility of EXE drug). The data of EXE solubility are used from a published source [23], and the experimental results were used to build, train, and assess the models. The list of EXE solubility is listed in Table 1, and the pairwise distribution of parameters of the process is illustrated in Fig. 1.

3. Computational methodology

In this research, the three methods introduced in the introduction section, which are tree-based ensemble methods, have been used as the core of the analysis for correlating the EXE solubility data as listed in Table 1. In general, we can summarize the modeling activity in the following items:

- Pre-Processing: data normalization and outlier detection.
- Model selection: hyper-parameter tanning in this research is done using Golden eagle optimizer (GEOA) [24]. GEOA is a meta-heuristic optimization methodology that offers benefits over conventional optimization methods. The algorithm begins with a basic population and then duplicates golden eagle hunting behavior and improves population fitness and discovers the optimal state. The golden eagle's instinct to soar and circle prey
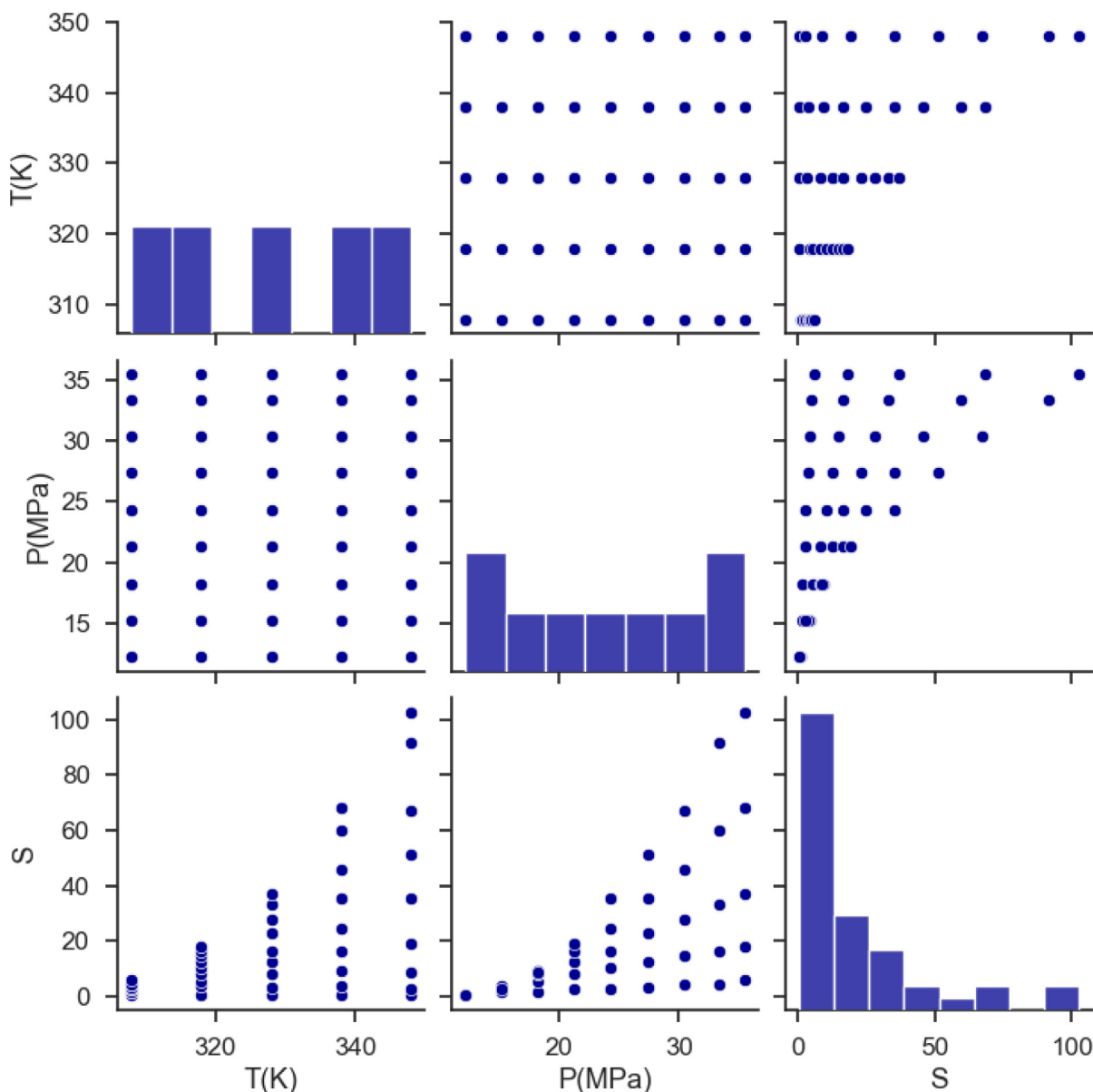
Fig. 1. Distributions of parameters for EXE solubility in Sc-CO$_2$.

serves as inspiration for GEOA. These golden eagles have excellent memories and will often tell other eagles where they last saw their prey. In order to address issues of exploitation and optimization, GEOA's numerical equations will model attack and cruise trajectories. The program takes a starting population and optimizes it by simulating the hunting strategies of the golden eagle to determine the most effective strategy. As a result, GEOA can also be employed to tackle real-world engineering challenges, where it excelled. The outputs indicate that GEOA is able to solve optimization issues with challenging and unknown search spaces and identify the global optimum [24,25].

- Evaluation: Finally, the optimized models were evaluated using visual and statistical methods.

3.1. Random forest and extra tree

Bagging methods such as RF and ET are useful to improve decision tree estimators. Random Forest (RF) is an ensemble learning model that incorporates voting to boost the efficiency of learners with numerous base trees [26]. A random forest's widespread acceptance stems from its ability to accurately forecast a wide range of outcomes with a small number of inputs. This technique effectively deals with both high-dimensional feature spaces and limited sample sizes. Since they may be run in parallel, they can manage realistically enormous systems [15]. The original dataset is bootstrapped into N instance sets in order to create an RF model. For each bootstrap sample, a fully-grown (unpruned) decision tree will be constructed. The following is the next step. During this stage, a collection of K base models is chosen at random to execute the function of dividing possibilities rather than using all available predictors. Iterations of C tree models with the aforementioned properties will be performed until the desired properties are obtained. Non-observed data will then be calculated by integrating the estimations from various C trees. Random Forest maximizes tree variety while minimizing model variance by constructing DTs from distinct training groups. The next equation shows an Random forest regression estimator formulation [27]:

Table 2
Final Models Errors.

Models	RMSE	MSE	MAPE
ET	1.522	2.317	0.2113
GB	2.351	5.531	0.4033
RF	2.827	7.991	0.3197

Table 3
Final Models R^2 score.

Models	Train R^2	Test R^2
ET	0.9945	0.9938
GB	0.9758	0.9853
RF	0.9904	0.9780
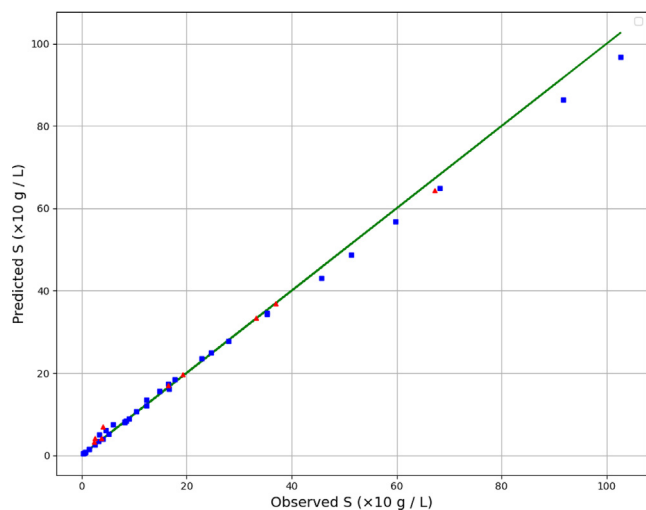
Fig. 2. Observed vs Predicted Values (ET Model).
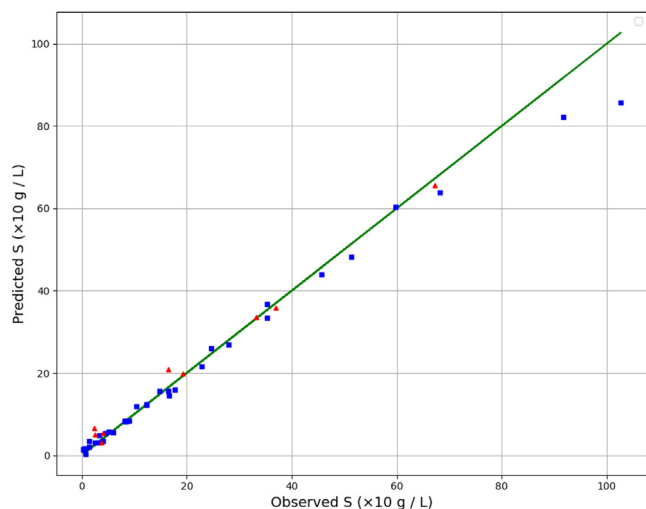
Fig. 3. Observed vs Predicted Values (GB Model).
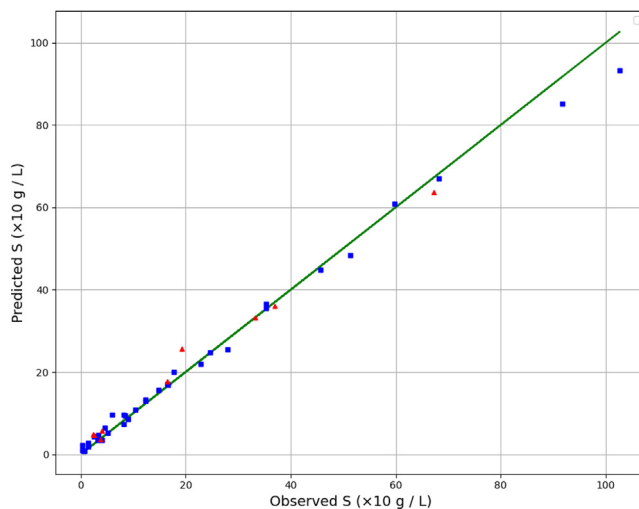
Fig. 4. Observed vs Predicted Values (RF Model).
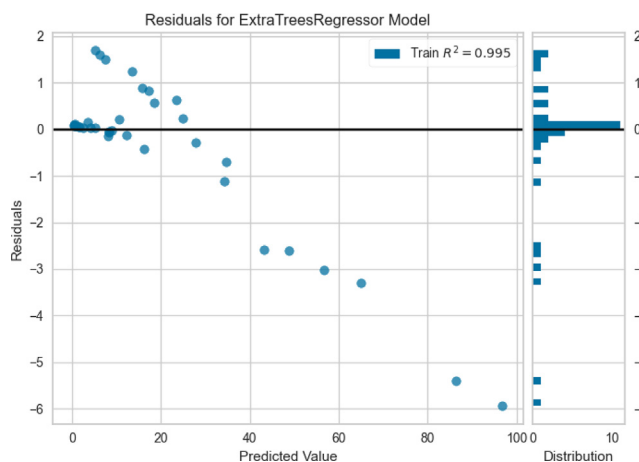
Fig. 5. Representation of the modelling Residuals of ET model.
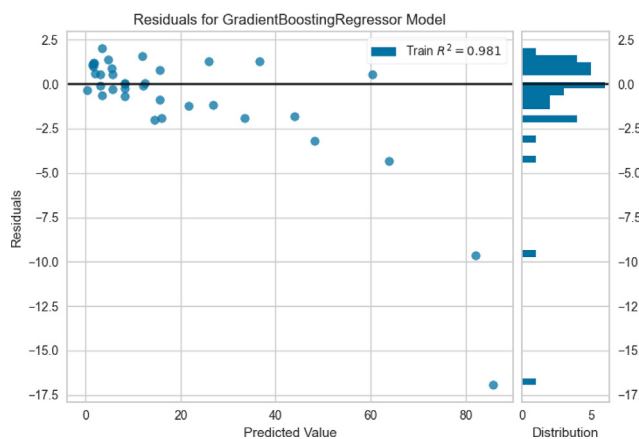
Fig. 6. Representation of the modelling Residuals of GB model.

$$\widehat{f}^{\,C}_{RF}(x) = \frac{1}{C}\sum_{i=1}^{C} T_i(x)$$

In this equation, C denotes the count of DTs, and x stands for the instance. $T_i(x)$ also reflects a unique Decision Tree based on bootstrap instances and a selection of entry variables. For the present being, it is possible to estimate out of bag (OOB) error using Ran-

dom Forest by evaluating instances that were not chosen in connection with the drive of this shaft during the bagging step. This sub-association will not utilize any external data to obtain an impartial estimate of generalization error [28,29]. Each input variable should be given a significant score. In this model, one input variable is changed while all other input variables remain constant, and the model's average decrease is also calculated [28,30].
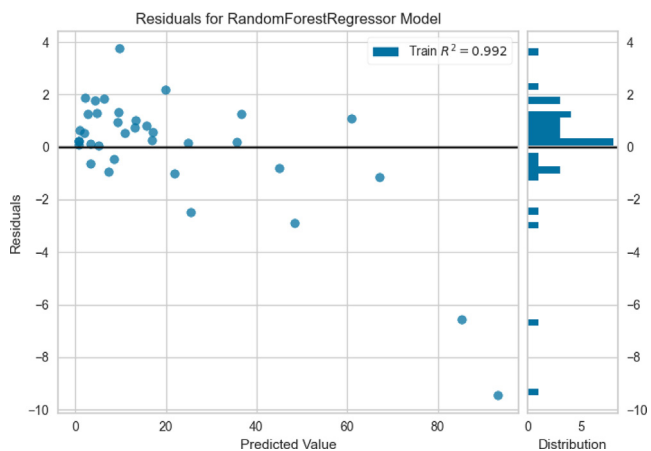
Fig. 7. Representation of the modeling residuals of RF model.

Extra Trees (ET), a tree-based method, are similar to random forests. ET must greatly randomize both the particularities of each tree node and the cut point choice during its division in order to categories and evaluate data in a way that is useful to the user [9,21,22].

Both methods are equivalent in terms of how they construct several trees and partition nodes using random subsets of functions. The main difference is that ET uses randomized divides rather than bootstrap data, rather than optimal splits [31].

3.2. Gradient boosting

The method of Gradient Boosting (GB) expands upon DT model by introducing a statistical approach whose fundamental premise is to employ a series of "weak" estimators to get a single "strong" ensemble estimator [32,33]. In GB, new decision trees are constructed successively by minimizing the current residuals. This
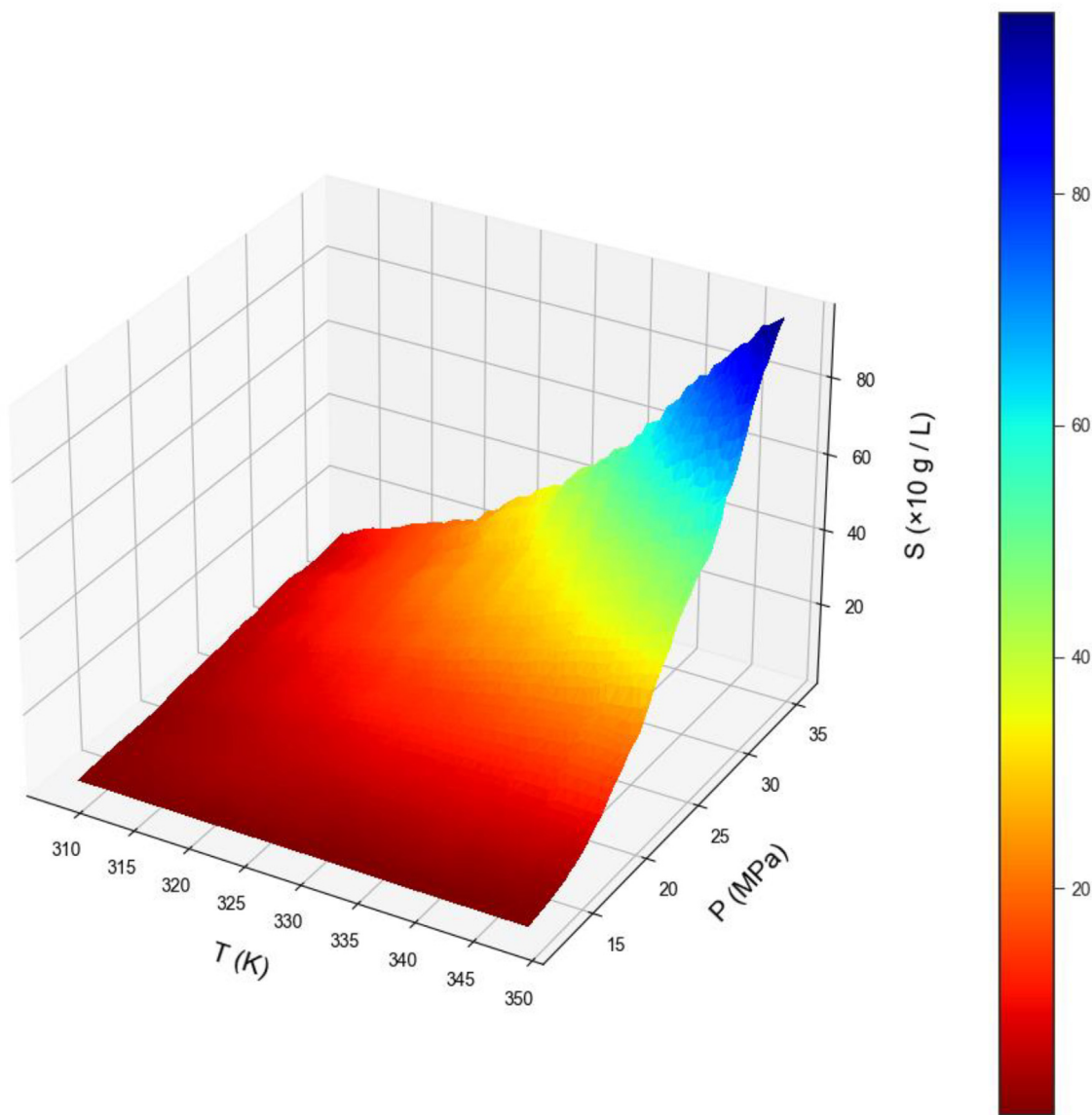

Fig. 8. Simulated prediction surface of solubility variations (ET model).

approach to continuous model generation is basically a type of functional gradient descent, in which estimate is maximized by creating a new base estimator (DT) to optimize the loss function at any stage [19,34,35]. The GB method is illustrated in the following algorithm:

Initialize $F_0(x) = argmin_p \sum_{i-1}^{N} L(y_i, P)$

For $k = 1 \, to \, M$:

   1. Determine the negative gradient

$$\bar{y}_i = -\left[\partial L \left(\frac{y_i, F(x_i)}{\partial F_{x_i}}\right)\right]$$ 2. Grow a DT model

$$a_k = argmin_{a,\beta} \sum_{i=1}^{N} \left[\bar{y} - \beta h(x_i, a_k)\right]^2$$ 3. Choose a gradient

descent step size

$$p_k = argmin_p \sum_{i=1}^{N} L(y_i, Fk - 1(x_i) + ph(x_i, a))$$ 4. Update

the approximation of F(x)

$$F_k(x) = F_k - 1(x) + p_k h(x, a_k)$$ Output: the final

estimation function $F_k(x)$

---

4. Results and discussion

Chosen ensemble models tuned using Golden Eagle Optimization Algorithm and final optimized estimators obtained. Then they implemented with their best configurations and evaluated. The error rates and R^2 scores of final models are displayed in Table 2 and Table 3.

Looking at Tables 2 and 3, the ET model is clearly better than the other two models. In addition to this analysis, Figs. 2–4 display the comparison of the experimentally observed values and the predicted values, in which the blue squares are the training data, and the red triangles are the test data. These three figures together show the fact that although all three models have high quality and efficiency, the ET model has the best result among them. The same can be seen for Figs. 5–7, which show the residuals of the models.

The model's representation as 3D and 2D plots are provided in Figs. 8–10 & Figs. 13–14, respectively, in which the predicted solubility surface is indicated versus temperature and pressure which are the two input variables of the machine learning models. It is clearly observed that the effect of *P* on the EXE solubility is more
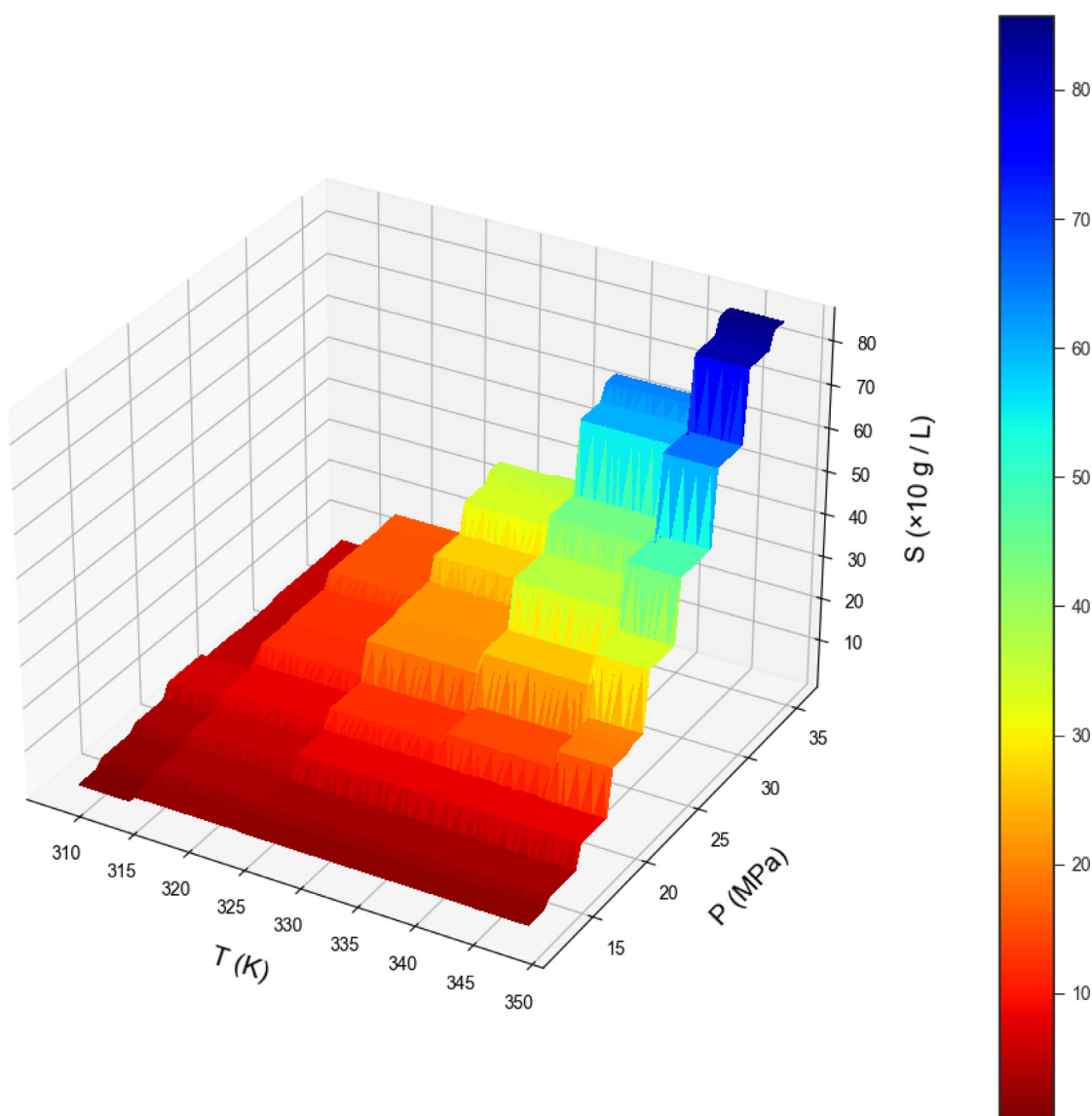

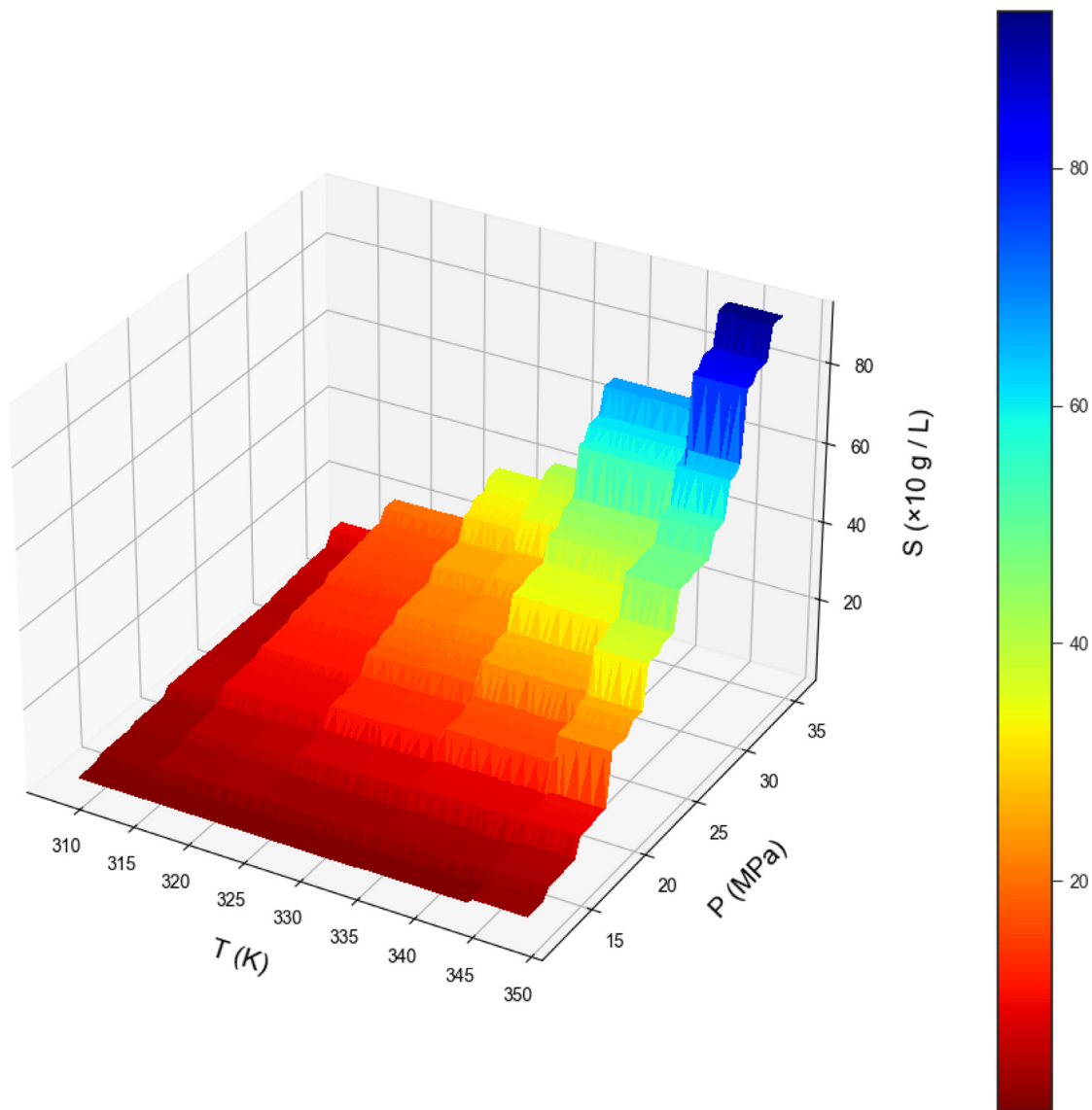**Fig. 9.** . Simulated prediction surface of solubility variations (GB model).

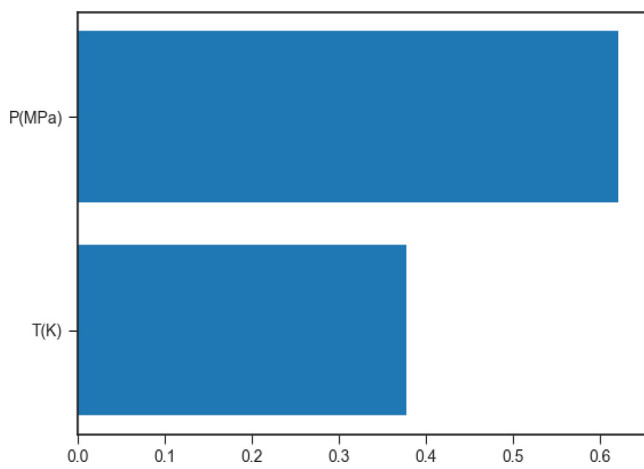**Fig. 10.** Simulated prediction surface of solubility variations (RF model).

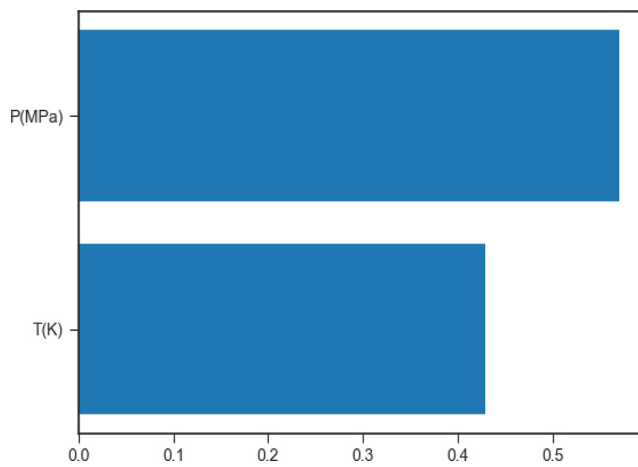**Fig. 11.** Feature Importance for solubility variations (using ET Model).

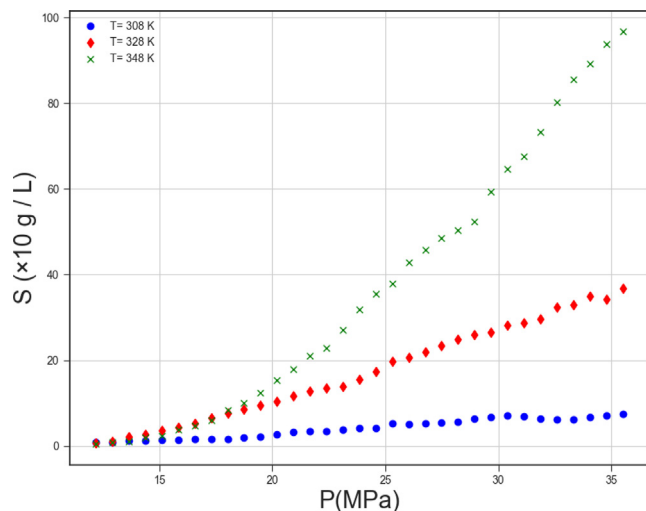**Fig. 12.** Feature Importance for solubility variations (using RF Model).

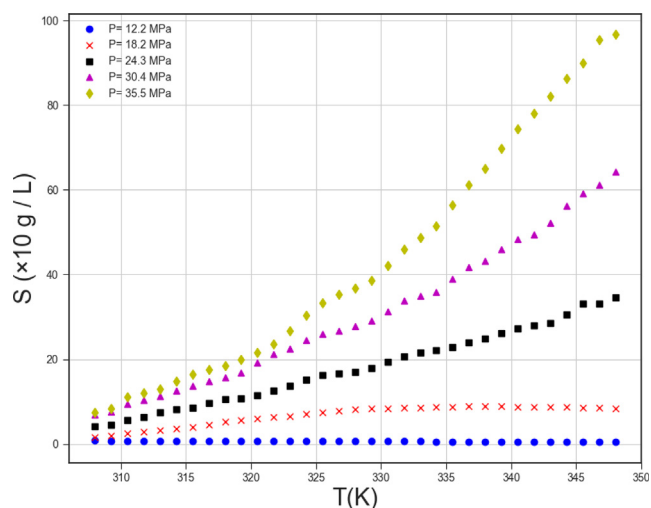**Fig. 13.** Trends of input parameter P on different Temperature levels.

Fig. 14. Trends of input parameter T on different Pressure levels.

recognized as the green method in pharmaceutical manufacturing. Using several tree-based ensemble approaches, we investigate how temperature and pressure affect EXE medication solubility. The Random Forest (RF), Extra Tree (ET), and Gradient Boosting models were used in this investigation (GB). As the main aspect of novelty of this research, Golden eagle optimizer (GEOA) is also employed for model optimization and hyper-parameter tweaking. The R^2 values for the ET, GB, and RF test phases were 0.993, 0.985, and 0.978, respectively. In train phases, the scores are 0.9945, 0.9758, and 0.9904. The ET model was specifically chosen since it is the most accurate and general. This model has error rates of 2.317 with MSE, 1.522 with RMSE, and 0.2113 with MAPE. The interpretation of the results indicated that pressure is more effective than the temperature on the solubility of EXE in the Sc-CO$_2$ as the green solvent.

Funding

CRediT authorship contribution statement

Bingfeng Luo: Conceptualization, Writing – original draft, Validation, Resources. **Tao Yang:** Validation, Supervision, Writing – review & editing, Investigation, Formal analysis. **Sabrean Farhan Jawad:** Software, Writing – review & editing, Data curation. **Hayder Imad Jabar:** Writing – review & editing, Formal analysis, Validation. **Hasan Khalid Dabis:** Writing – original draft, Validation, Software. **Mohaned Adil:** Writing – original draft, Validation, Resources. **Anfal Nabeel Mustafa:** Writing – review & editing, Validation, Investigation. **Salema K. Hadrawi:** Writing – review & editing, Formal analysis, Resources. **Ibrahim Mourad Mohammed:** Conceptualization, Writing – original draft. **Abdullah Alshetaili:** Conceptualization, Data curation, Validation. **Naseer Mehdi Mohammed:** Data curation, Validation, Supervision. **Umme Hani:** Funding acquisition, Software, Formal analysis. **Amal M. Alsubaiyel:** Software, Formal analysis, Resources.

Data availability

All data are available within the published paper.

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.

Acknowledgments

References

[1] W.K. Abdelbasset et al., Development a novel robust method to enhance the solubility of Oxaprozin as nonsteroidal anti-inflammatory drug based on machine-learning, Sci. Rep. 12 (1) (2022) 13138.

[2] M.A.S. Abourehab et al., Theoretical investigations on the manufacture of drug nanoparticles using green supercritical processing: estimation and prediction of drug solubility in the solvent using advanced methods, J. Mol. Liq. (2022).

[3] L.A. Blanchard, J.F. Brennecke, Recovery of organic products from ionic liquids using supercritical carbon dioxide, Ind. Eng. Chem. Res. 40 (1) (2001) 287–292.

profound than the effect of *T* which could be attributed to the fact that the density of the solvent changes more with the pressure change as it is in the gas state. It is also seen that T has attractive effects on the EXE solubility, so that both factors have positive effects in the EXE solubility values. It can be perceived that both pressure and temperature must be enhanced in order to increase the solubility of the drug in the solvent for preparation of the nanoparticles.

Also, one of the unique features of bagging models based on decision trees is the ease of finding the importance of features. Accordingly, in Figs. 11 and 12, the importance of two features is shown with the final models of ET and RF, and in both of them, the pressure feature is more important than the temperature which is consistent with the physical justifications.

5. Conclusion

Computational simulation of pharmaceutical solubility in supercritical solvent was carried out in this study via an advanced hybrid modeling approach. The process of supercritical has been

[4] H. Chinh Nguyen et al., Computational prediction of drug solubility in supercritical carbon dioxide: thermodynamic and artificial intelligence modeling, J. Mol. Liq. 354 (2022).

[5] F. An et al., Machine learning model for prediction of drug solubility in supercritical solvent: modeling and experimental validation, J. Mol. Liq. 363 (2022).

[6] H. Bagheri, G. Ali Mansoori, H. Hashemipour, A novel approach to predict drugs solubility in supercritical solvents for RESS process using various cubic EoS-mixing rule, J. Mol. Liq. 261 (2018) 174–188.

[7] K.D. Bartle et al., Solubilities of solids and liquids of low volatility in supercritical carbon dioxide, J. Phys. Chem. Ref. Data 20 (4) (1991) 713–756.

[8] R.B. Chim et al., Solubility of dexamethasone in supercritical carbon dioxide, J. Chem. Eng. Data 57 (12) (2012) 3756–3760.

[9] Y. Li et al., Theoretical modeling study on preparation of nanosized drugs using supercritical-based processing: determination of solubility of Chlorothiazide in Supercritical Carbon dioxide, J. Mol. Liq. (2022).

[10] A. Alobaida, B. Huwaimel, Analysis of enhancing drug bioavailability via nanomedicine production approach using green chemistry route: systematic assessment of drug candidacy, J. Mol. Liq. (2022).

[11] S.M. Alshahrani et al., Green processing based on supercritical carbon dioxide for preparation of nanomedicine: model development using machine learning and experimental validation, Case Stud. Therm. Eng. 41 (2023).

[12] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2019.

[13] S. González et al., A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities, Information Fusion 64 (2020) 205–237.

[14] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm. in ICML, Citeseer, 1996.

[15] R. Seyghaly et al., Interference recognition for fog enabled IoT architecture using a novel tree-based method, In: 2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS), IEEE Computer Society, 2022.

[16] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[17] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE, 1995.

[18] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[19] L. Mason et al., Boosting algorithms as gradient descent, Adv. Neural Inf. Proces. Syst. 12 (1999).

[20] A. Verikas, A. Gelzinis, M. Bacauskiene, Mining data with random forests: a survey and results of new tests, Pattern Recogn. 44 (2) (2011) 330–349.

[21] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.

[22] S. Dutta, U. Mukherjee, S.K. Bandyopadhyay, Pharmacy impact on Covid-19 vaccination progress using machine learning approach, J. Pharm. Res. Int. (2021) 202–217.

[23] M. Hojjati et al., Supercritical CO_2 and highly selective aromatase inhibitors: experimental solubility and empirical data correlation, J. Supercrit. Fluids 50 (3) (2009) 203–209.

[24] A. Mohammadi-Balani et al., Golden eagle optimizer: a nature-inspired metaheuristic algorithm, Comput. Ind. Eng. 152 (2021).

[25] T. Magesh, G. Devi, T. Lakshmanan, Improving the performance of grid connected wind generator with a PI control scheme based on the metaheuristic golden eagle optimization algorithm, Electr. Pow. Syst. Res. 214 (2023).

[26] R. Jiang et al., A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinf. 10 (1) (2009) 1–12.

[27] V. Rodriguez-Galiano 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 (2015) 804–818.

[28] L. Breiman et al., Classification and Regression Trees, Routledge, 2017.

[29] J. Zhang et al., Rapid evaluation of texture parameters of Tan mutton using hyperspectral imaging with optimization algorithms, Food Control (2022).

[30] S. Alshehri et al., Design of predictive model to optimize the solubility of Oxaprozin as nonsteroidal anti-inflammatory drug, Sci. Rep. 12 (1) (2022) 13106.

[31] Y.-Y. Song, L. Ying, Decision tree methods: applications for classification and prediction, Shanghai Arch. Psychiatry 27 (2) (2015) 130.

[32] H.-T. Lin, T.-J. Liang, S.-M. Chen, Estimation of battery state of health using probabilistic neural network, IEEE Trans. Ind. Inf. 9 (2) (2012) 679–685.

[33] H. Li et al., Optimization and design of machine learning computational technique for prediction of physical separation process, Arab. J. Chem. 15 (4) (2022).

[34] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813.

[35] V.-H. Truong et al., A robust method for safety evaluation of steel trusses using Gradient Tree Boosting algorithm, Adv. Eng. Softw. 147 (2020).