Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Novel and robust machine learning approach for estimating the fouling factor in heat exchangers

Saleh Hosseini [a], Amith Khandakar [b,*], Muhammad E.H. Chowdhury [b]
Mohamed Arselene Ayari [c,d], Tawsifur Rahman [b], Moajjem Hossain Chowdhury [e],
Behzad Vaferi [f,*]

[a] Department of Chemical Engineering, University of Larestan, Larestan, Iran
[b] Electrical Engineering Department, Qatar University, Doha 2713, Qatar
[c] Department of Civil and Architectural Engineering, College of Engineering, Qatar University, Doha 2713, Qatar
[d] Technology Innovation and Engineering Education, College of Engineering, Qatar University, Doha 2713, Qatar
[e] Department of Electrical, Electronics and Systems Engineering, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[f] Department of Advanced Calculations, Chemical, Petroleum, and Polymer Engineering Research Center, Shiraz Branch, Islamic Azad University, Shiraz, Iran

ARTICLE INFO

Article history:
Received 31 December 2021
Received in revised form 16 May 2022
Accepted 29 June 2022
Available online 6 July 2022

Keywords:
Heat exchange
Fouling
Gaussian Process Regression
Decision trees
Support Vector Regression
Comparison study

ABSTRACT

The fouling factor (R_f) is an operating index for measuring an undesirable effect of solids' deposition on the heat transfer ability of heat exchangers. Accurate prediction of the fouling factor helps appropriate scheduling of the cleaning cycles. Since diverse factors affect this operating feature, it is sometimes hard to estimate the fouling factor accurately using simple empirical or traditional intelligent methods. Therefore, this study employs four up-to-date machine-learning algorithms (Gaussian Process Regression, Decision Trees, Bagged Trees, Support Vector Regression) and a traditional model (Linear Regression) to estimate the fouling factor as a function of operating and constructing variables. The 5-fold cross-validation using 9268 data samples determines the structure of the considered estimators, and 2358 external datasets have been utilized for models' testing. The relevancy analysis confirms that the most accurate predictions are achieved when the square root of the fouling factor ($\sqrt{R_f}$) is simulated. The Gaussian Process Regression (GPR) shows the highest level of agreement with the experimental samples in both the model construction and testing stages. The trained GPR model scored an R^2 value of 0.98770 and 0.99857 on the internal and external datasets, respectively. The model predicts the overall 11626 experimental samples (Davoudi and Vaferi, 2018) with the MAPE = 13.89%, MSE = 7.02×10^{-4}, and R^2 = 0.98999. The proposed GPR model outperforms the previously suggested artificial neural network for estimating the fouling factor in heat exchangers.

1. Introduction

Process monitoring is a common practice to ensure product quantity (Qiao et al., 2022) and quality (Fortuna et al., 2005), equipment health (Musa et al., 2021; Said et al., 2022), minimize hazardous materials and potential risks (Bernechea and Viger, 2013), enhance energy efficiency (Ejaz et al., 2021; Mota et al., 2018), reduce loss level (Hosseini and Vaferi, 2022), and so on. Heat exchangers are among the most widely used thermal-based equipment in diverse domestic and industrial applications, including HVAC (heating, ventilation, and air conditioning) (Sukarno et al., 2021; Sun et al., 2021), bone micro-grinding (Yang et al., 2021), food processing (Dekhil et al., 2020), oil (Lai et al., 2011), gas refinery (Fard et al., 2017), reforming process (Barnoon, 2021), petrochemical complex (Brodowicz and Markowski, 2003). This energy-based equipment transfers heat between several operating and utility streams with different temperatures (Holman, 2010). They have an essential role in effectively utilizing available energy sources (Ma et al., 2016) and reducing utility costs (Ravagnani et al., 2005).

During heat exchangers' operations, heat transfer surface areas are possibly covered with various deposits available in either fluid streams or formed by chemical/biological reactions and corrosion inside a system (Cui et al., 2021; Davoudi and Vaferi, 2018; Holman, 2010). These deposits may have different sources, such as microbial (Xu et al., 2016), mineral (Kazi et al., 2015), corrosion (Ren et al., 2019), particulate (Inamdar et al., 2016), and chemical reaction (Ishiyama et al., 2017). Coating the heat transfer surface

* Corresponding authors.
E-mail addresses: amitk@qu.edu.qa (A. Khandakar), behzad.vaferi@iau.ac.ir (B. Vaferi).

https://doi.org/10.1016/j.egyr.2022.06.123

area by deposited solid particles increases thermal resistance and decreases the efficiency of heat exchangers (Davoudi and Vaferi, 2018). Although the fouling factor or resistance (R_f) is widely employed as a parameter to describe how fouling affects a heat exchanger (Genić et al., 2012; Holman, 2010), the literature has also highlighted several uncertainties associated with the fouling resistance (Al-Janabi et al., 2009; Sloley, 2021). The overall heat-transfer coefficients of dirty (U_{dirty}) and clean (U_{clean}) surfaces determined by experimental measurements of temperatures and thermal flux are often used to calculate the R_f by Eq. (1) (Holman, 2010).

$$R_f = U_{dirty}^{-1} - U_{clean}^{-1} \tag{1}$$

The R_f unit is m^2 K/kW. The fouling resistance behavior is a fundamental parameter for determining the cleaning schedule (Lozano Santamaria and Macchietto, 2019). The interested readers are better refer to Lozano Santamaria and Macchietto study to better familiar with the cleaning schedule of heat exchangers (Lozano Santamaria and Macchietto, 2019). This type of cleaning stage decreases processing time, increases plant downtime, and its economic cost (Wallhäuzer et al., 2013).

There are different scenarios to detect/monitor the fouling factor in heat exchangers. These methods rely on monitoring the weight of clean and fouled surfaces (Tissier and Lalande, 1986), flow rates and temperatures (Kuwahara et al., 2015; Liporace and De Oliveira, 2007; Polley et al., 2007), overall heat transfer coefficient (Shen et al., 2018), wall shear stress (Genić et al., 2012), and electrical (Chen et al., 2004), photothermal (Fujimori et al., 1987), and acoustic and optical (Withers, 1996) signals to measure the fouling factor in heat exchangers. Since relatively all the laboratory and real-field measurements are often affected by different uncertainty sources, some researchers employed computational paradigms to estimate the fouling factor in heat exchangers (Yang, 2020).

The main advantage of the computational methods is that they only use some previously measured variables to calculate the fouling factor (Davoudi and Vaferi, 2018), and their online version can be easily updated during process operation (Gudmundsson et al., 2016). Diverse mathematical and intelligent methods such as feedforward (Lalot and Pálsson, 2010), recurrent (Shaosheng and Ju, 2007), Chebyshev (Fan and Zhong, 2013), NARX (nonlinear autoregressive network with exogenous inputs) (Kumari and Srinivasan, 2017), and deep (Sundar et al., 2020) neural networks, least-squares (Ying and Nan, 2010) and conventional support vector machines (Sun et al., 2008), k-means (Wang and Fan, 2012) and genetic (Tang et al., 2020) algorithms, extended Kalman filter (Jonsson et al., 2007), computational fluid dynamics (Bayat et al., 2012), polynomial fuzzy (Delmotte et al., 2013), and fuzzy Takagi–Sugeno representation (Delrot et al., 2012), and long short-term memory (Madhu PK et al., 2021) have been used to either estimate or detect fouling phenomenon in heat exchangers. All these proposed techniques are either developed by small numbers of actual information or they provide a relatively high level of prediction errors. Therefore, their generalization or/and prediction accuracy are under doubt.

In the current research, 11626 actual samples previously analyzed by Davoudi and Vaferi (Davoudi and Vaferi, 2018) have been used to develop an up-to-date and efficient smart methodology for accurately calculating the fouling factor in the heat exchanger. For doing so, three novel approaches (Gaussian Process Regression, Decision Trees, and Bagged Trees) along with two traditional ones (Support Vector Regression and Linear Regression) have been considered. Furthermore, an efficient 5-fold cross-validation with the relevancy analysis is also engaged in the current research. A systematic uncertainty monitoring and ranking analysis helped choose the highest accurate model for the given objective. To the best of the authors' knowledge, the constructed Gaussian Process Regression (GPR) in this work is not only the most robust model (developed by 11626 actual samples out of which 9268 data samples were used to train the model and the trained model was tested on external 2358 data samples), but it also presents the most accurate predictions.

Table 1
Number (percent) of samples in the internal and external datasets.

Internal dataset		External dataset
Training	Validation	Testing
7414 (63.77%)	1854 (15.95%)	2358 (20.28%)

2. Methodology

In this section, the experiment data is explained. The machine learning models used will be described in detail, along with the model development process and the measurement criteria utilized to evaluate the performance of considered models.

2.1. Models' development process

The complete methodology performed in the current study to estimate the fouling factor in heat exchangers is schematically presented in Fig. 1. A part of the 11626 collected experimental samples (see Section 3), termed internal data, is engaged in the 5-fold cross-validation to choose the best topology of the considered machine learning models as well as find the most efficient R_f transformation. The performance of the constructed machines is then tested on a completely unseen dataset known as the external dataset.

The 5-fold cross-validation splits the internal database into five equal folds first. Then one-fold is held out, and the model is trained on the other four folds. The fold that is not used validates the performance of the model. This is continued until all folds are engaged in the validation stage. In this way, all five folds engage in both training and validation stages. This allows the construction of more robust and generalizable intelligent estimators.

The number and percentage of the internal (training and validation) and external (testing) databases are reported in Table 1. It can be seen that eighty percent of the available samples are used during the 5-fold cross-validation to identify the topology of the machines and determine the R_f transformation. The remaining twenty unseen samples have been utilized to check the generalization ability of the constructed estimators.

2.2. Machine learning algorithm

In this study, four different machine learning models (Gaussian Process Regression, Decision Trees, Bagged Trees, Support Vector Regression) and a traditional technique (i.e., Linear Regression) have been employed for this regression task. The underlying mechanisms behind these algorithms have been explained in this subsection.

2.2.1. Gaussian process regression (GPR)

GPR is a Bayesian regression approach that works well on small datasets (Zazoum, 2021). Where most the supervised machine learning algorithms learn the exact values of the function for each parameter, GPR learns a distribution of probability over all possible values (Williams and Rasmussen, 2006).

A linear function is assumed to explain this (Eq. (2)).

$$y = wX + \varepsilon \tag{2}$$

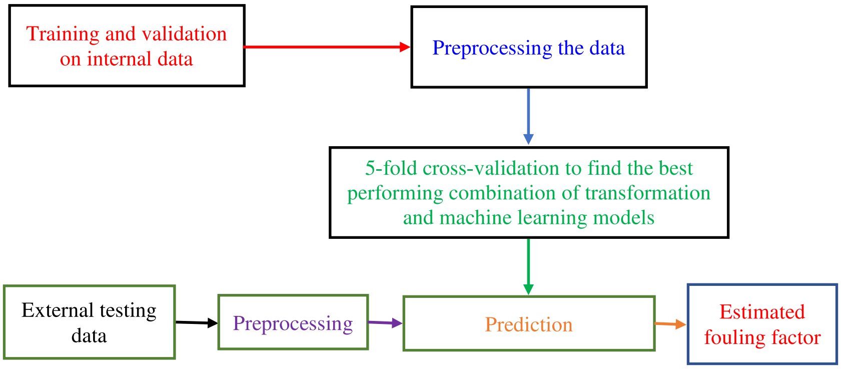
Training and validation on internal data
Preprocessing the data
5-fold cross-validation to find the best performing combination of transformation and machine learning models
External testing data
Preprocessing
Prediction
Estimated fouling factor

Fig. 1. Flowchart of smart models' development and validation.

GPR will work by first taking a prior distribution, $p(w)$, on the parameter, w, and shifting the probabilities based on the training data using Bayes' rule. The posterior distribution can be calculated as follows:

$$Posterior = \frac{likelihood \times prior}{marginal\ likelihood} \tag{3}$$

$$p\left(w|y, X\right) = \frac{p\left(y|X, w\right) \times p\left(w\right)}{p\left(y|X\right)} \tag{4}$$

The posterior distribution is calculated to incorporate meaningful information from both the training data and prior distribution into it.

2.2.2. Decision trees

Decision Trees are a non-parametric supervised learning method (Naik and Ianakiev, 2021). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Khosravi et al., 2018). A tree can be seen as a piecewise constant approximation.

2.2.3. Bagged trees

Bootstrap aggregation (bagging) is a type of ensemble learning (Hussain et al., 2021). To bag a weak learner such as a decision tree on a dataset, generate many bootstrap replicas of the dataset and grow decision trees on the replicas. Obtain each bootstrap replica by randomly selecting N out of N observations with replacement, where N is the dataset size. In addition, every tree in the ensemble can randomly select predictors for each decision split, and a technique called random forest (Breiman, 2001, 1996) is known to improve the accuracy of bagged trees.

2.2.4. Support vector regression (SVR)

SVR is the support vector machine used for regression tasks (Ighravwe and Mashao, 2020). The SVR algorithm was first proposed by Vapnik in 2013 (Vapnik, 2013). SVR is non-parametric as it uses kernel functions to build the model (Fan et al., 2005). SVR minimizes L1 Loss or mean absolute error as shown below:

$$L1Loss = \left(\frac{1}{n}\right)\sum_{k=1}^{n}|y_k - \widehat{y_k}| \tag{5}$$

where y represents the ground truth, $\hat{y}$ represents the prediction, and n shows the number of samples. The model essentially maps the features such that each prediction deviates from its ground truth by a value not larger than ε (default is 0.1 in Matlab) but also ensures that the decision surface is as flat as possible.

2.2.5. Linear regression

Linear regression is a basic type of predictive modeling that explicitly explains the relationship between a dependent and independent variable (Bagherzadeh et al., 2022). The model created has linear coefficients. Linear regression minimizes the residual sum of squares between the ground truth and predictions. This type of modeling, while easy to understand, is not robust.

2.3. Models' accuracy measurement

Monitoring the prediction accuracy of a given model is an important stage for determining its most efficient topology and comparing its performance with another potential/available scenarios. Diverse statistical uncertainty criteria are available to monitor the deviation between actual data and their related predictions. This study considers mean squared errors (MSE), mean absolute error (MAE), regression coefficient (R^2-value), relative absolute error (RAE%), and mean absolute percentage error (MAPE%). The mathematical expressions of these criteria are defined *by* Eq. (6) to Eq. (10) (Cao et al., 2022; Qiao et al., 2021).

$$MAPE\% = (100/n) \times \sum_{k=1}^{n}\left|\hat{y} - y\right| / y \tag{6}$$

$$MSE = (1/n)\sum_{k=1}^{n}\left(\hat{y} - y\right)^2 \tag{7}$$

$$MAE = (1/n) \times \sum_{k=1}^{n}\left|\hat{y} - y\right| \tag{8}$$

$$RAE\% = 100 \times \sum_{k=1}^{n}\left|\hat{y} - y\right| / \sum_{k=1}^{n}|y - \bar{y}| \tag{9}$$

$$R^2 = 1 - \left[\sum_{k=1}^{n}\left(\hat{y} - y\right)^2 / \sum_{k=1}^{n}\left(y - \bar{y}\right)^2\right] \tag{10}$$

here $\bar{y}$ is the mean of the ground truths, and n represents the number of samples.

3. Laboratory-measured fouling samples

As mentioned earlier, the current work utilizes 11626 laboratory-measured fouling factors for developing as well as testing the considered methods. The complete information about this database has been presented in the supplementary materials. The fouling factor has been reported as a function of fluid chemistry (composition) and velocity, its oxygen content, surface and fluid temperature, operation time, and available cross-section for fluid flow.

3.1. Preprocessing

The present section reviews normalization (scaling) and R_f transformation accomplished as two main preprocessing stages.

3.1.1. Scaling (normalization)

Before beginning the training stage of a machine-learning model, it is important to normalize all features to a similar range. This is because features with different magnitudes distort the training stage and increase its computational effort (Zhao et al., 2021). In the worst case, it results in an optimization algorithm in local minima (with poor performance). The features were normalized using Z-score normalization, as shown in Eq. (11).

$$\bar{x} = \frac{(x-\mu)}{\sigma} \tag{11}$$

where x, μ, and σ are the original, average, and standard deviation values of the feature, and $\bar{x}$ is the normalized feature. This process ensures that the average and standard deviation of the normalized features are zero and one, respectively.

3.1.2. Transformation

Pearson's method is an efficient tool for monitoring the dependency of a dependent variable on its independent variables (Hosseini and Vaferi, 2022). As Eq. (12) shows, Pearson's method measures the interrelation between a pair of dependent–independent (τ, γ) variables by a $\eta_{\tau\gamma}$ factor (Karimi et al., 2020). The highest (i.e., +1) and smallest (i.e., -1) factors indicate the strongest direct and indirect dependency of dependent to independent variables, respectively. Zero or close value to zero means there is no relationship between the variables.

$$\eta_{\tau\gamma} = \sum_{k=1}^{n} (\tau_k - \bar{\tau})(\gamma_k - \bar{\gamma}) \Big/ \left(\sqrt{\sum_{k=1}^{n} (\tau_k - \bar{\tau})^2} \sqrt{\sum_{k=1}^{n} (\gamma_k - \bar{\gamma})^2} \right) \tag{12}$$

here, $\bar{\tau}$ (Eq. (13)) and $\bar{\gamma}$ (Eq. (14)) are average values of independent and dependent variables, respectively.

$$\bar{\tau} = (1/n) \times \sum_{k=1}^{n} \tau_k \tag{13}$$

$$\bar{\gamma} = (1/n) \times \sum_{k=1}^{n} \gamma_k \tag{14}$$

Therefore, it can be claimed that diverse transformations of the fouling factor in heat exchangers can provide different generalization abilities during the modeling stage. This issue will be accomplished in Section 4.2. The prediction accuracy of the considered estimators over various transformations of the fouling factor will be monitored, and the best one will be chosen.

3.2. Relevancy examination

The consequences of applying Pearson's method to the available experimental fouling database are displayed in Fig. 2a. This figure implies that the fouling factor directly relates to the operation time, surface temperature, and fluid velocity. This observation states that the fouling factor grows by increasing the surface temperature, operation time, and fluid velocity. On the other hand, the negative Pearson's coefficients for the fluid density, fluid temperature, and equivalent diameter indicate an indirect relationship between the fouling factor and these features. The highest positive and negative coefficients indicate that operation time and equivalent diameter are the features with the strongest direct and indirect effects on the fouling factor.

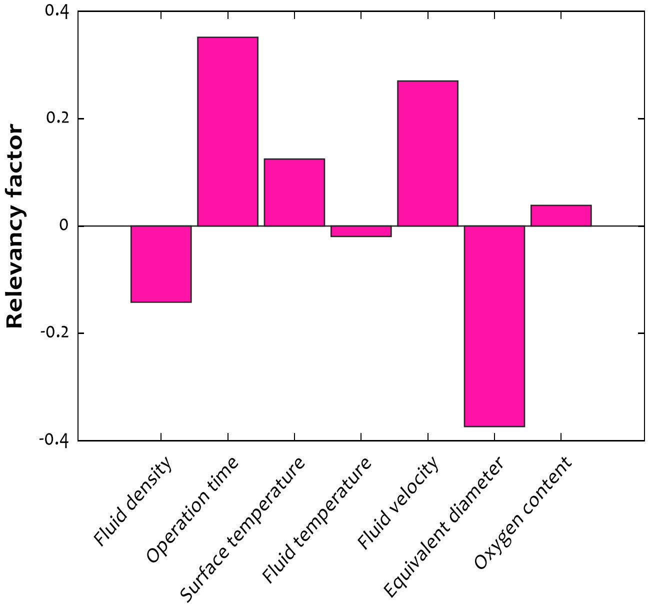

Fig. 2a. Relevancy factors between fouling factor and the corresponding features determined by the Pearson method.

It is better to note that the anticipation of Pearson's method conflicts with the literature observation for the effect of fluid velocity and equivalent diameter on the fouling factor. The literature correctly states that increasing the fluid velocity and/or decreasing the equivalent diameter decreases the fouling rate (Asomaning, 1997; Sundaram, 1998; Wilson et al., 2017). Based on Eq. (12), Pearson's method only monitors the variation of the dependent variable (γ) by only one independent variable (τ). Indeed, this statistical-based method ignores the combined effect of other variables on the dependent variable.

Figs. 2b and 2c exhibit the variation of fouling factor by the fluid velocity and equivalent diameter, respectively. These figures also present the linear equations fitted to the experimental data. It can be seen that these linear equations (analogous to Pearson's method) wrongly show that the fouling factor increases by increasing the velocity and decreasing the equivalent diameter. The highly scattering of the experimental data may also be responsible for this wrong anticipation of Pearson's method.

4. Results and discussion

This section presents the detailed findings and provides some explanations for justifying the obtained results.

4.1. Modeling the fouling factor

As mentioned earlier, ~80% of the fouling measurement samples have been used in the 5-fold cross-validation procedure to regulate the structural features of the considered estimators. The R^2 values provided by these five machine-learning models are reported in Table 2. It can be seen that GPR, bagged trees, and decision trees exhibit the best accuracy performance in terms of R^2-value.

It should be noted that the previous models have been constructed using the original values of the fouling factor, and no transformation has been considered here before. The effect of the non-linear transformation of the fouling factor on the prediction performances of the considered estimators is investigated in the next section.

Table 2
The observed R^2 for predicting the R_f values in the internal datasets by different smart models.

Smart estimator	GPR	Decision trees	Bagged trees	SVR	Linear regression
R^2	0.9754	0.9686	0.9712	0.9222	0.7523

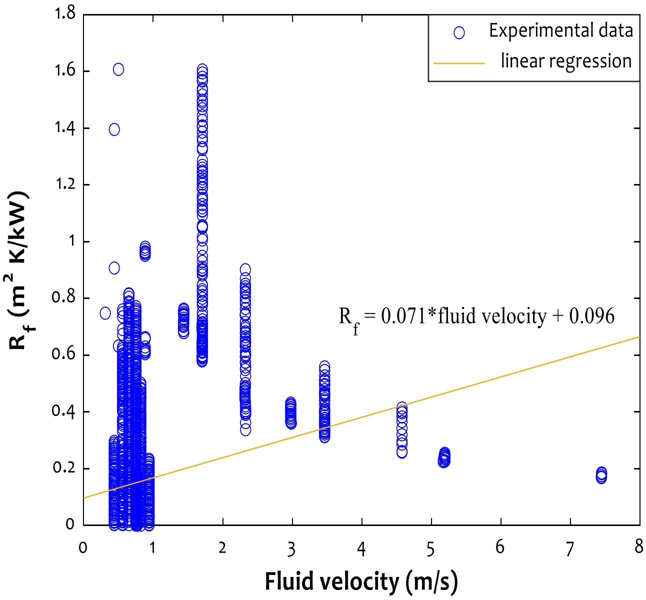

Fig. 2b. Variation of the fouling factor by the fluid velocity (without considering the effect of other parameters) and its associated linear regression.

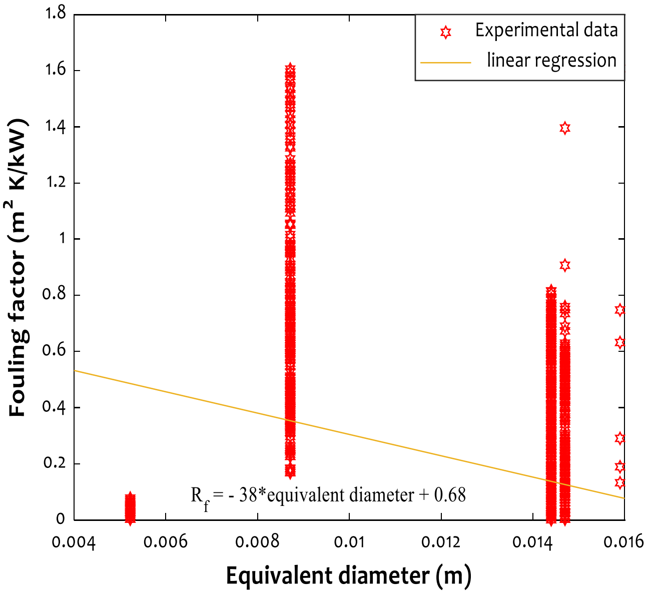

Fig. 2c. Variation of the fouling factor by the equivalent diameter (without considering the effect of other parameters) and its associated linear regression.

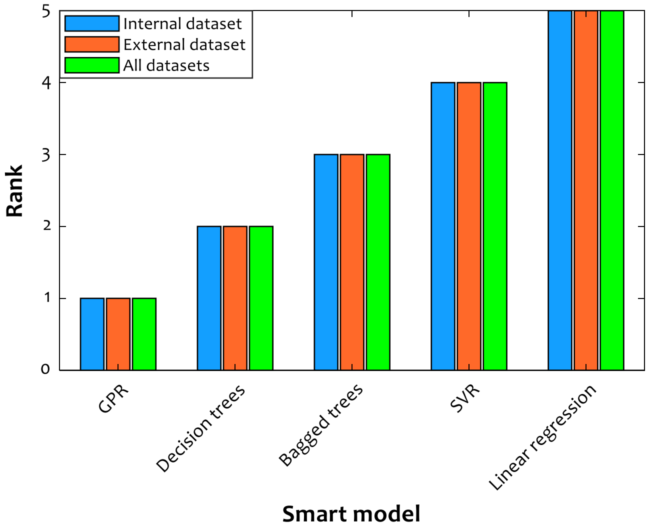

Fig. 3. The outcome of applying ranking analysis to the reported results in Table 4.

4.2. Modeling the transformed fouling factor

It was previously justified (see Section 3.1.2) that the R_f transformation may positively affect the generalization ability of the considered intelligent estimators. Transforming the ground truth often helps capture the highest possible relationship between the predictors and the response variable (fouling factor). This section studies the influence of the non-linear transformation of the R_f on the observed R^2-values during a 5-fold cross-validation process (Table 3). The results reported in Table 3 approve that $\sqrt{R_f}$ gave the best results for all the employed models. The observed results indicate that the $\sqrt{R_f}$ is better estimated using the considered tools. Indeed, the modeling phase focuses on simulating the $\sqrt{R_f}$ variable. But, all analyses have been done based on the R_f values (it means that the applied transformation should be reversed to reach the R_f value).

Table 4 shows the prediction performance of the GPR, decision trees, bagged trees, SVR, and linear regression models during the 5-fold cross-validation step and a testing stage for estimating the R_f. Five statistical criteria (i.e., MAPE%, MAE, RAE%, MSE, and R^2-value) have been applied to monitor the prediction performance of these models.

It is easy to conclude that the SVR and linear regression models provided worse results than the GPR, decision, and bagged trees paradigms. On the other hand, it is hard to find the most accurate method only using visual observation. Therefore, Section 4.3 employs a systematic ranking analysis to order the constructed models based on their performances measured by five statistical indices.

4.3. The most accurate model for estimating the $\sqrt{R_f}$

The previous section utilized five statistical indices to measure the prediction accuracy of the fabricated GPR, decision trees, bagged trees, SVR, and linear regression for the R_f estimation. The ranking analyses have been performed on the accuracy of these models over the internal, external, and complete databases (Fig. 3). The obtained results by the ranking analysis approve that the GPR has the best performance in the 5-fold cross-validation and testing stages. Therefore, this intelligent model achieves the first overall ranking. In addition, the linear regression with the fifth ranking places over the internal and external databases is the worst model for estimating the R_f experimental samples.

Supplementary materials report the GPR predictions for each data point in the internal and external databases.

Table 3
Cross-validation results for estimating different non-linear transformations of the fouling factor.

Smart estimator	The observed R^2 for the different R_f transformations						
	R_f^2	R_f	$\sqrt{R_f}$	$\sqrt[4]{R_f}$	$R_f^{0.1}$	log (R_f)	exp (R_f)
GPR	0.8677	0.9754	**0.9844**	0.9841	0.9741	0.9621	0.9497
Decision trees	0.9230	0.9686	**0.9785**	0.9762	0.9647	0.9480	0.9402
Bagged trees	0.9076	0.9712	**0.9802**	0.9756	0.9659	0.9495	0.9401
SVR	0.5323	0.9222	**0.9659**	0.9531	0.9336	0.9089	0.7843
Linear regression	0.5902	0.7523	**0.7731**	0.7418	0.7028	0.6577	0.6872

Table 4
Monitoring the accuracy of the developed machine learning methods for R_f estimating using five statistical criteria.

Smart estimator	Database	MAPE%	MAE	RAE (%)	MSE	R^2
GPR	Internal	13.55	5.05×10^{-3}	3.97	8.53×10^{-4}	0.98770
	External	15.20	3.75×10^{-3}	2.87	1.06×10^{-4}	0.99857
	All data	13.89	4.78×10^{-3}	3.74	7.02×10^{-4}	0.98999
Decision trees	Internal	16.14	8.84×10^{-3}	6.95	9.22×10^{-4}	0.98664
	External	18.61	6.07×10^{-3}	4.65	1.91×10^{-4}	0.99736
	All data	16.64	8.27×10^{-3}	6.47	7.73×10^{-4}	0.98890
Bagged trees	Internal	28.64	1.22×10^{-2}	9.62	1.15×10^{-3}	0.98484
	External	29.70	1.11×10^{-2}	8.53	5.59×10^{-4}	0.99358
	All data	28.86	1.20×10^{-2}	9.39	1.03×10^{-3}	0.98668
SVR	Internal	35.36	1.50×10^{-2}	11.81	1.65×10^{-3}	0.97702
	External	35.48	1.40×10^{-2}	10.71	7.86×10^{-4}	0.98982
	All data	35.39	1.48×10^{-2}	11.59	1.48×10^{-3}	0.97971
Linear regression	Internal	80.89	4.65×10^{-2}	36.55	4.98×10^{-2}	0.57753
	External	88.45	6.27×10^{-2}	48.01	9.37×10^{-1}	0.19943
	All data	82.42	4.98×10^{-2}	38.92	2.30×10^{-1}	0.32546

Table 5
Comparison of the accuracy of the GPR and literature proposed MLPNN for calculating the R_f value.

	MAPE%	MAE	RAE (%)	MSE	R^2
MLPNN	33.15	2.01×10^{-2}	15.70	1.88×10^{-3}	0.94634
GPR	13.89	4.78×10^{-3}	3.74	7.02×10^{-4}	0.98999
Improvement	138.7	319.6	319.6	168.3	4.4

4.4. Comparison with literature

This section compares the prediction accuracy of the GPR model with the previously proposed intelligent model in the literature. Davoudi and Vaferi constructed a conventional multilayer perceptron neural network (MLPNN) using the same database utilized in the current study (Davoudi and Vaferi, 2018). Their MLPNN model provided the most accurate results for predicting the fourth roots of the fouling factor ($\sqrt[4]{R_f}$). For comparison on an identical basis, the GPR and MLPNN accuracy for estimating the original R_f values has been considered. Table 5 presents the result of this comparison in terms of MAPE%, MAE, RAE%, MSE, and R^2. The reported indices in Table 5 approved that the accuracy of the newly developed GPR is far better than the conventional MLPNN. The last row of Table 5 quantizes the improvement level achieved by the GPR against the MLPNN. This table shows that the improvements range from 4.4% for R^2 to 319.6% for MAE and RAE%. It is necessary to highlight that Eq. (15) is used to calculate the improvement percent.

$$\%Improvement = 100 \times |MLPNN - GPR| / MLPNN \tag{15}$$

4.5. Performance evaluation of the GPR model

Here before, it is justified that the fabricated GPR has the highest accuracy for predicting huge numbers of experimental fouling samples in heat exchangers. This section investigates the performance of the constructed GPR using the visual inspection (cross-plot, residual error, Kernel density estimation, and Bland–Altman plot).

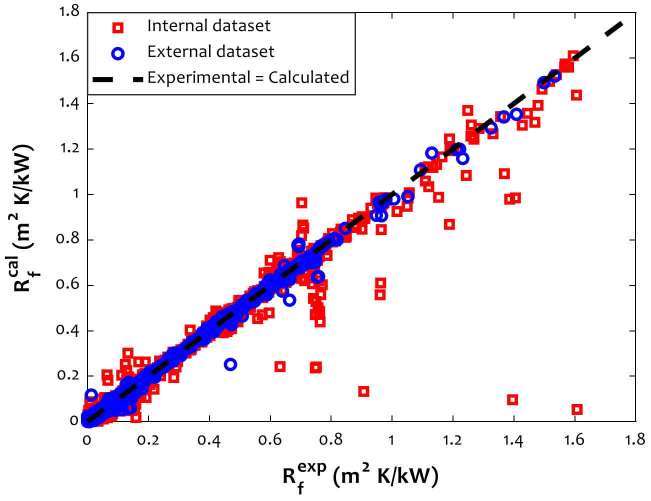

Fig. 4. Regression plot of the GPR predictions for the R_f against their ground truths for the internal and external datasets.

Fig. 4 shows the regression plot for the GPR models on the internal and external databases. The idea behind the regression plot is that with a perfect model, the regression line (the line fitted to the scatter plot) will be the same as the dotted diagonal line. That is easy to see that every prediction will be the same as the ground truth. Observing Fig. 4 shows that most of the predictions are very close to that ideal line, and indeed the regression line itself almost coincides with the $y = x$ line. This is further evident because the correlation coefficients between output and target are 0.98770 and 0.99857 for 5-fold cross-validation and testing stages, respectively. Hence it can be said that the model achieved an almost perfect result.

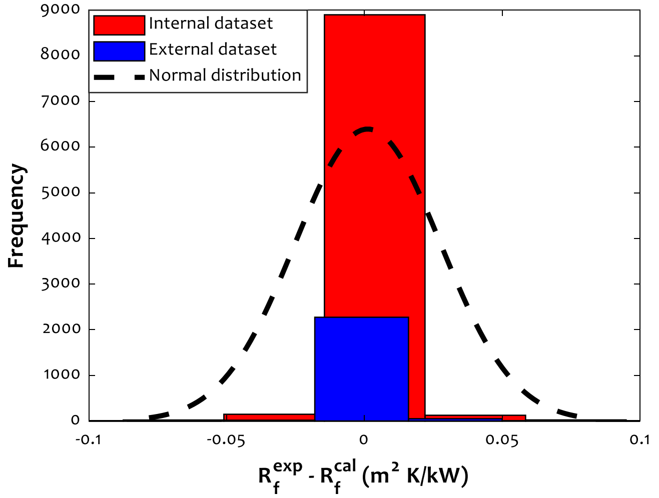

Fig. 5. The observed histogram for the residual error for estimating the R_f (Standard deviation and average values are 0.02646 and 0.00122 m^2 K/kW).

Eqs. (16) to (18) have been applied to measure the residual error (RE) between experimental and calculated values of the R_f, its average ($\overline{RE}$) and standard deviation (SD).

$$RE_k = \left(R_f^{\exp} - R_f^{cal}\right)_k \quad k = 1, 2, \ldots, n \tag{16}$$

$$\overline{RE} = \frac{1}{n} \times \sum_{k=1}^{n} RE_k \tag{17}$$

$$SD = \sqrt{\sum_{k=1}^{n} \left(RE - \overline{RE}\right)_k^2 / n} \tag{18}$$

Fig. 5 depicts the histogram of the monitored RE between the experimental and calculated values of the R_f. This figure indicates that the RE covers an infinitesimal domain from −0.05 to +0.05. Furthermore, the highest number of internal and external samples have been estimated with negligible RE value (∼0).

Bland–Altman plot shows the spread of the predicted data samples by a given model (Karvaly et al., 2017). Furthermore, the 95% limit of agreement can be seen from the plot. A smaller limit of agreement means a better model which would be indicated by a very tight grouping. Figs. 6a and 6b illustrate the Bland–Altman plot for the GPR predictions on the internal as well as external databanks, respectively. The dashed lines are associated with the upper and lower limits of agreement (LoA). Numerical values of the upper and lower LoA can be reached using Eqs. (19) and (20), respectively.

$$Upper\ LoA = \overline{RE} + 1.96 \times SD \tag{19}$$

$$Lower\ LoA = \overline{RE} - 1.96 \times SD \tag{20}$$

The data points are very tightly grouped and are very close to 0. The 95% limit of agreement for the internal databank (Fig. 6a) is from −0.05583 to 0.05857, and for the external database (Fig. 6b) is from −0.01953 to 0.02082. This shows that the ensembled GPR model could handle all data types and was not biased to a specific group.

The distribution of the predictions by the ensembled GPR model and the ground truth values for the internal and external databanks obtained by the Kernel density estimation (Lacour et al., 2017) is depicted in Figs. 7a and 7b, respectively. It can be observed that the prediction distribution was nearly identical to the distribution of the ground truth. The distributions are slightly different between 0.6 and 0.8. Hence, these plots justify the robustness of the ensemble GPR model.

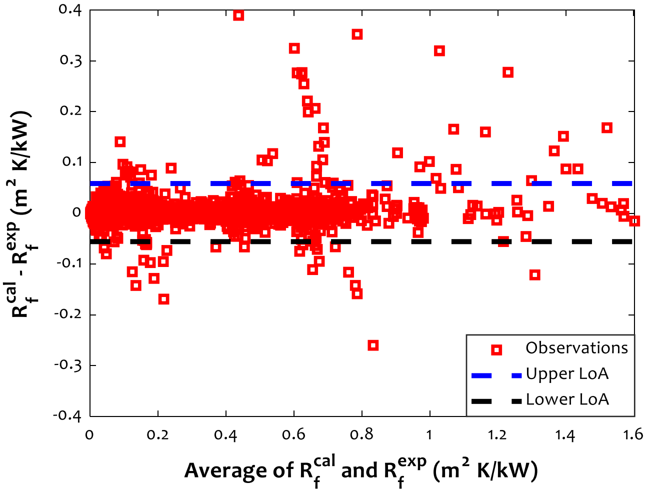

Fig. 6a. Bland–Altman plot for the internal group.

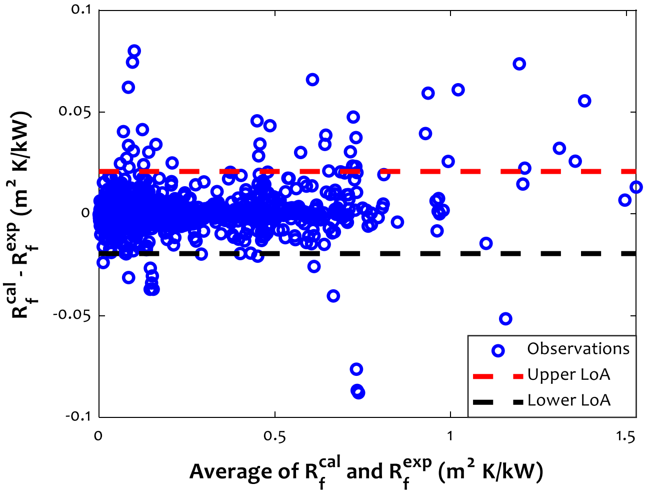

Fig. 6b. Bland–Altman plot for the external group.

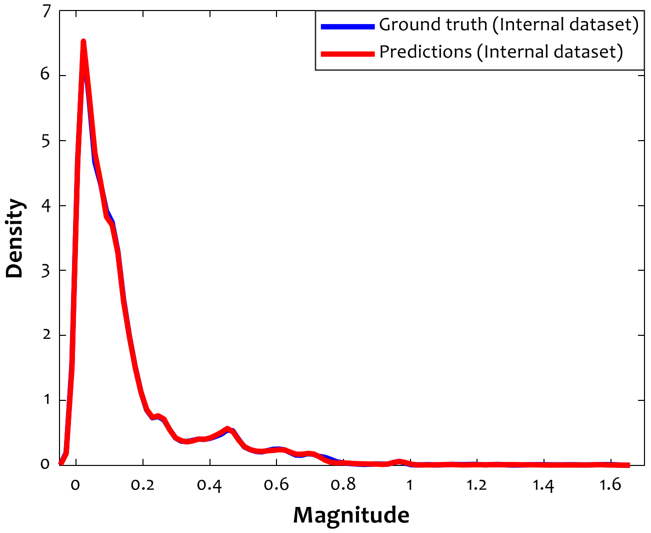

Fig. 7a. Kernel density estimation (KDE) plot for the internal datasets.

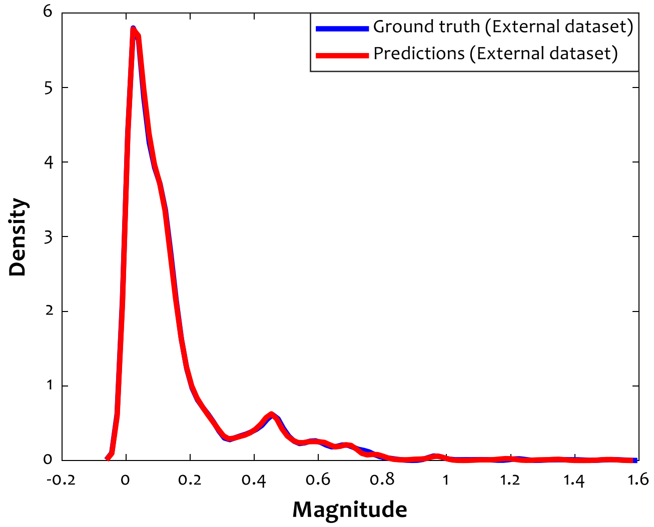

Fig. 7b. Kernel density estimation plot for the external datasets.

4.6. Potential application of the GPR model

The selected intelligent paradigm by the ranking analyses (i.e., Gaussian Process Regression) estimates the fouling factor accurately when the conditions are within the applicability realm considered (the source of database) or close to that involved in this study.

In addition, the GPR paradigm can be applied with care for estimating the fouling factor in those cases that do not cover the above criterion, as its results may be less accurate in these cases. However, as the GPR model was constructed using a comprehensive fouling database covering many operating conditions, it has broad applicability potential. Therefore, it can be considered for fouling monitoring in many situations.

5. Conclusion

A machine learning-based approach has been deployed in this study to predict the fouling factor in heat exchangers under a wide range of operating conditions. Seven non-linear transformations have been applied to the fouling factor, and tha accuracy of five different methods (four machine-learning models and a traditional approach) have been compared over the transformations. $\sqrt{R_f}$ is empirically found to give the best result. The best model (GPR) has been tested on both an internal dataset (5-fold cross-validation) and an external dataset. The current state-of-the-art model provides an R^2 of 0.98999 compared to the previous best result of 0.94634. Hence, the model has proven to be both robust and accurate in estimating the fouling factor. Because classical machine learning models inherently have fewer parameters than big neural networks, this model can be easily deployed to an edge device. The GPR model accurately estimates the fouling factor and helps associated industries during the maintenance of heat exchangers.

CRediT authorship contribution statement

Saleh Hosseini: Drafting of the actual manuscript and also in the rebuttal preparation, Involved in formal investigation. **Amith Khandakar:** Drafting of the actual manuscript and also in the rebuttal preparation, Worked on developing the trained network. **Muhammad E.H. Chowdhury:** Drafting of the actual manuscript and also in the rebuttal preparation, Involved in formal investigation. **Mohamed Arselene Ayari:** Drafting of the actual manuscript and also in the rebuttal preparation, Involved in formal investigation. **Tawsifur Rahman:** Drafting of the actual manuscript and also in the rebuttal preparation, Worked on developing the trained network. **Moajjem Hossain Chowdhury:** Drafting of the actual manuscript and also in the rebuttal preparation, Worked on developing the trained network. **Behzad Vaferi:** Drafting of the actual manuscript and also in the rebuttal preparation, Involved in formal investigation.

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.

Acknowledgment

The publication of this article was funded by the Qatar National Library.

Appendix A. Supplementary data

All collected experimental data and GPR predictions for each individual data point of the internal and external data point have been reported in the supplementary excel files. Furthermore, a user-friendly code has been provided/prepared to utilize by the research community.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.egyr.2022.06.123.

References

Al-Janabi, A., Esawy, M., Malayeri, M.R., Müller-Steinhagen, H., 2009. Consideration of dynamic uncertainty in fouling experimentation. In: Proc. of International Conference on Heat Exchanger Fouling and Cleaning VIII-June. pp. 14–19.

Asomaning, S., 1997. Heat Exchanger Fouling By Petroleum Asphaltenes. University of British Columbia Vancouver, BC, Canada.

Bagherzadeh, A., Shahini, N., Saber, D., Yousefi, P., Seyed Alizadeh, S.M., Ahmadi, S., Tat Shahdost, F., 2022. Developing a global approach for determining the molar heat capacity of deep eutectic solvents. Meas. J. Int. Meas. Confed 188, 110630.

Barnoon, P., 2021. Modeling of a high temperature heat exchanger to supply hydrogen required by fuel cells through reforming process. Energy Rep. 7, 5685–5699.

Bayat, M., Aminian, J., Bazmi, M., Shahhosseini, S., Sharifi, K., 2012. CFD modeling of fouling in crude oil pre-heaters. Energy Convers. Manag 64, 344–350.

Bernechea, E.J., Viger, J.A., 2013. Design optimization of hazardous substance storage facilities to minimize project risk. Saf. Sci. 51, 49–62.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brodowicz, K., Markowski, M., 2003. Calculation of heat exchanger networks for limiting fouling effects in the petrochemical industry. Appl. Therm. Eng. 23, 2241–2253.

Cao, Y., Kamrani, E., Mirzaei, S., Khandakar, A., Vaferi, B., 2022. Electrical efficiency of the photovoltaic/thermal collectors cooled by nanofluids: Machine learning simulation and optimization by evolutionary algorithm. Energy Rep. 8, 24–36.

Chen, X.D., Li, D.X.Y., Lin, S.X.Q., Özkan, N., 2004. Online fouling/cleaning detection by measuring electric resistance-equipment development and application to milk fouling detection and chemical cleaning monitoring. J. Food Eng. 61, 181–189.

Cui, H., Ye, B., Wang, L., Li, N., Xing, D., 2021. Reconciling thermal performance and power-saving performance of counter-flow spray heating towers. Energy Rep. 7, 1529–1538.

Davoudi, E., Vaferi, B., 2018. Applying artificial neural networks for systematic estimation of degree of fouling in heat exchangers. Chem. Eng. Res. Des. 130, 138–153.

Dekhil, M.A., Tala, J.V.S., Bulliard-Sauret, O., Bougeard, D., 2020. Development of an innovative heat exchanger for sensible heat storage in agro-food industry. Appl. Therm. Eng. 177, 115412.

Delmotte, F., Dambrine, M., Delrot, S., Lalot, S., 2013. Fouling detection in a heat exchanger: A polynomial fuzzy observer approach. Control Eng. Pract. 21, 1386–1395.

Delrot, S., Guerra, T.M., Dambrine, M., Delmotte, F., 2012. Fouling detection in a heat exchanger by observer of takagi–sugeno type for systems with unknown polynomial inputs. Eng. Appl. Artif. Intell. 25, 1558–1566.

Ejaz, A., Babar, H., Ali, H.M., Jamil, F., Janjua, M.M., Fattah, I.M.R., Said, Z., Li, C., 2021. Concentrated photovoltaics as light harvesters: Outlook, recent progress, and challenges. Sustain. Energy Technol. Assess. 46, 101199.

Fan, R.E., Chen, P.H., Lin, C.J., Joachims, T., 2005. Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. 6, 1889–1918.

Fan, S., Zhong, Q., 2013. Prediction of fouling in condenser based on fuzzy stage identification and chebyshev neural network. Meas. Sci. Rev. 13, 94–99.

Fard, M.M., Pourfayaz, F., Kasaeian, A.B., Mehrpooya, M., 2017. A practical approach to heat exchanger network design in a complex natural gas refinery. J. Nat. Gas Sci. Eng. 40, 141–158.

Fortuna, L., Graziani, S., Xibilia, M.G., 2005. Soft sensors for product quality monitoring in debutanizer distillation columns. Control Eng. Pract. 13, 499–508.

Fujimori, H., Asakura, Y., Suzuki, K., Uchida, S., 1987. Noncontact measurement of film thickness by the photothermal deflection method. Jpn. J. Appl. Phys. 26, 1759–1764.

Genić, S.B., Jaćimović, .B.M., Mandić, D., Petrović, D., 2012. Experimental determination of fouling factor on plate heat exchangers in district heating system. Energy Build. 50, 204–211.

Gudmundsson, O., Palsson, O.P., Palsson, H., Lalot, S., 2016. Online fouling detection of domestic hot water heat exchangers. Heat Transf. Eng. 37, 1231–1241.

Holman, J.P., 2010. Heat Transfer. McGraw-Hill, Boston, USA.

Hosseini, S., Vaferi, B., 2022. Determination of methanol loss due to vaporization in gas hydrate inhibition process using intelligent connectionist paradigms. Arab. J. Sci. Eng. 47, 5811–5819. http://dx.doi.org/10.1007/s13369-021-05679-4.

Hussain, S., Mustafa, M.W., Jumani, T.A., Baloch, S.K., Alotaibi, H., Khan, I., Khan, A., 2021. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Rep. 7, 4425–4436.

Ighravwe, D.E., Mashao, D., 2020. Analysis of support vector regression kernels for energy storage efficiency prediction. Energy Rep. 6, 634–639.

Inamdar, H.V, Groll, E.A., Weibel, J.A., Garimella, S.V, 2016. Prediction of air-side particulate fouling of HVAC & R heat exchangers. Appl. Therm. Eng. 104, 720–733.

Ishiyama, E.M., Kennedy, J., Pugh, S.J., 2017. Fouling management of thermal cracking units. Heat Transf. Eng. 38, 694–702.

Jonsson, G.R., Lalot, S., Palsson, O.P., Desmet, B., 2007. Use of extended Kalman filtering in detecting fouling in heat exchangers. Int. J. Heat Mass Transf. 50, 2643–2655.

Karimi, M., Vaferi, B., Hosseini, S.H., Olazar, M., Rashidi, S., 2020. Smart computing approach for design and scale-up of conical spouted beds with open-sided draft tubes. Particuology 55, 179–190.

Karvaly, G., Mészáros, K., Kovács, K., Patócs, A., Sipák, Z., Vásárhelyi, B., 2017. Looking beyond linear regression and bland-altman plots: A comparison of the clinical performance of 25-hydroxyvitamin D tests. Clin. Chem. Lab. Med. 55, 385–393.

Kazi, S.N., Teng, K.H., Zakaria, M.S., Sadeghinezhad, E., Bakar, M.A., 2015. Study of mineral fouling mitigation on heat exchanger surface. Desalination 367, 248–254.

Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755.

Kumari, S.A., Srinivasan, S., 2017. Narx models for prediction of reheater fouling. In: 2017 23rd International Conference on Automation and Computing. ICAC, IEEE, pp. 1–4.

Kuwahara, T., Wibowo, C., Kuboyama, A., Nakamura, M., Yamane, Y., 2015. Fouling monitoring in thermosiphon reboiler. Heat Transf. Eng. 36, 780–786.

Lacour, C., Massart, P., Rivoirard, V., 2017. Estimator selection: A new method with applications to kernel density estimation. Sankhya A 79, 298–335.

Lai, S.M., Wu, H., Hui, C.W., Hua, B., Zhang, G., 2011. Flexible heat exchanger network design for low-temperature heat utilization in oil refinery. Asia-Pacific. J. Chem. Eng. 6, 713–733.

Lalot, S., Pálsson, H., 2010. Detection of fouling in a cross-flow heat exchanger using a neural network based technique. Int. J. Therm. Sci. 49, 675–679.

Liporace, F. dos S., De Oliveira, S.G., 2007. Real time fouling diagnosis and heat exchanger performance. Heat Transf. Eng. 28, 193–201.

Lozano Santamaria, F., Macchietto, S., 2019. Online integration of optimal cleaning scheduling and control of heat exchanger networks under fouling. Ind. Eng. Chem. Res. 59, 2471–2490.

Ma, H., Yin, L., Shen, X., Lu, W., Sun, Y., Zhang, Y., Deng, N., 2016. Experimental study on heat pipe assisted heat exchanger used for industrial waste heat recovery. Appl. Energy 169, 177–186.

Madhu PK, R., Subbaiah, J., Krithivasan, K., 2021. RF-LSTM-based method for prediction and diagnosis of fouling in heat exchanger. Asia-Pacific. J. Chem. Eng. 16, e2684.

Mota, L., Mota, A., Pezzuto, C., Carvalho, M., Lavorato, M., Coiado, L., Oliveira, E., 2018. Development of a surface temperature sensor to enhance energy efficiency actions in buildings. Sensors 18 (3046).

Musa, G., Alrashed, M., Muhammad, N.M., 2021. Development of big data lean optimisation using different control mode for gas turbine engine health monitoring. Energy Rep. 7, 4872–4881.

Naik, K., Ianakiev, A., 2021. Heat demand prediction: A real-life data model vs simulated data model comparison. Energy Rep. 7, 380–388.

Polley, G.T., Wilson, D.I., Pugh, S.J., Petitjean, E., 2007. Extraction of crude oil fouling model parameters from plant exchanger monitoring. Heat Transf. Eng. 28, 185–192.

Qiao, W., Li, Z., Liu, W., Liu, E., 2022. Fastest-growing source prediction of US electricity production based on a novel hybrid model using wavelet transform. Int. J. Energy Res. 46, 1766–1788.

Qiao, W., Wang, Y., Zhang, J., Tian, W., Tian, Y., Yang, Q., 2021. An innovative coupled model in view of wavelet transform for predicting short-term PM10 concentration. J. Environ. Manage 289, 112438.

Ravagnani, M., Silva, A.P., Arroyo, P.A., Constantino, A.A., 2005. Heat exchanger network synthesis and optimisation using genetic algorithm. Appl. Therm. Eng. 25, 1003–1017.

Ren, L., Cheng, Y., Feng, S., Han, Z., 2019. Experimental study on corrosion-fouling relationship of Ni-WP composite coating surface of heat exchanger. Surf. Topogr. Metrol. Prop. 7 (15011).

Said, Z., Arora, S., Farooq, S., Sundar, L.S., Li, C., Allouhi, A., 2022. Recent advances on improved optical, thermal, and radiative characteristics of plasmonic nanofluids: Academic insights and perspectives. Sol. Energy Mater. Sol. Cells 236, 111504. http://dx.doi.org/10.1016/j.solmat.2021.111504.

Shaosheng, F., Ju, W., 2007. Application of diagonal recurrent neural network for measuring fouling in condenser. In: 2007 2nd IEEE Conference on Industrial Electronics and Applications. IEEE, pp. 169–171.

Shen, C., Wang, Y., Zhao, Z., Jiang, Y., Yao, Y., 2018. Decoupling analysis on the variations of liquid velocity and heat flux in the test of fouling thermal resistance. Int. J. Heat Mass Transf. 123, 227–238.

Sloley, A., 2021. Heat exchanger: Forestall fouling-factor foul-ups. [WWW Document]. URL https://www.chemicalprocessing.com/articles/2021/heat-exchanger-forestall-fouling-factor-foul-ups/. (Accessed 5 September 2022).

Sukarno, R., Putra, N., Hakim, I.I., Rachman, F.F., Mahlia, T.M.I., 2021. Utilizing heat pipe heat exchanger to reduce the energy consumption of airborne infection isolation hospital room HVAC system. J. Build. Eng. 35, 102116.

Sun, F., Yu, J., Zhao, A., Zhou, M., 2021. Optimizing multi-chiller dispatch in HVAC system using equilibrium optimization algorithm. Energy Rep. 7, 5997–6013.

Sun, L., Zhang, Y., Zheng, X., Yang, S., Qin, Y., 2008. Research on the fouling prediction of heat exchanger based on support vector machine. In: 2008 International Conference on Intelligent Computation Technology and Automation. ICICTA, IEEE, pp. 240–244.

Sundar, S., Rajagopal, M.C., Zhao, H., Kuntumalla, G., Meng, Y., Chang, H.C., Shao, C., Ferreira, P., Miljkovic, N., Sinha, S., 2020. Fouling modeling and prediction approach for heat exchangers using deep learning. Int. J. Heat Mass Transf. 159, 120112.

Sundaram, B.N., 1998. The effects of oxygen on synthetic crude oil fouling. Department of Chemical and Bio-Resource Engineering. University of British Columbia, Vancouver.

Tang, S.Z., Li, M.J., Wang, F.L., He, Y.L., Tao, W.Q., 2020. Fouling potential prediction and multi-objective optimization of a flue gas heat exchanger using neural networks and genetic algorithms. Int. J. Heat Mass Transf. 152, 119488.

Tissier, J.P., Lalande, M., 1986. Experimental device and methods for studying milk deposit formation on heat exchange surfaces. Biotechnol. Prog. 2, 218–229.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer science & business media.

Wallhäuzer, E., Hussein, M.A., Becker, T., 2013. A5. 4-clean or not clean-detecting fouling in heat exchangers. In: Proceedings SENSOR 2013. pp. 121–125.

Wang, S., Fan, S., 2012. Prediction of fouling in condenser based on k-means algorithms and improved Chebyshev neural network. In: International Conference on Automatic Control and Artificial Intelligence. ACAI 2012, IET, pp. 1596–1600.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning. MIT press Cambridge, MA.

Wilson, D.I., Ishiyama, E.M., Polley, G.T., 2017. Twenty years of Ebert and Panchal—What next? Heat Transf. Eng. 38, 669–680.

Withers, P.M., 1996. Ultrasonic, acoustic and optical techniques for the non-invasive detection of fouling in food processing equipment. Trends Food Sci. Technol. 7, 293–298.

Xu, Z., Wang, J., Jia, Y., Geng, X., Liu, Z., 2016. Experimental study on microbial fouling characteristics of the plate heat exchanger. Appl. Therm. Eng. 108, 150–157.

Yang, J., 2020. Computational fluid dynamics studies on the induction period of crude oil fouling in a heat exchanger tube. Int. J. Heat Mass Transf. 159, 120129.

Yang, M., Li, C., Luo, L., Li, R., Long, Y., 2021. Predictive model of convective heat transfer coefficient in bone micro-grinding using nanofluid aerosol cooling. Int. Commun. Heat Mass Transf. 125, 105317.

Ying, Z., Nan, W.Y., 2010. Prediction of condenser fouling based on locally weighted partial least squares regression algorithm. Chin. J. Sci. Instrum. 2 (13).

Zazoum, B., 2021. Solar photovoltaic power prediction using different machine learning methods. Energy Rep..

Zhao, T.H., Khan, M.I., Chu, Y.M., 2021. Artificial neural networking (ANN) analysis for heat and entropy generation in flow of non-Newtonian fluid between two rotating disks. Math. Methods Appl. Sci. 1–19. http://dx.doi.org/10.1002/mma.7310.