Journal of Environmental Management 292 (2021) 112736

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Developing machine learning models for relative humidity prediction in air-based energy systems and environmental management applications

Kinza Qadeer [a], Ashfaq Ahmad [b], Muhammad Abdul Qyyum [a,**], Abdul-Sattar Nizami [c], Moonyong Lee [a,*]

[a] School of Chemical Engineering, Yeungnam University, Gyeongsan, 712-749, Republic of Korea
[b] Department of Computer Science, COMSATS University Islamabad (CUI), Lahore Campus, Defense Road, Off Raiwind Road, Lahore, Pakistan
[c] Sustainable Development Study Center, Government College University, Lahore, 54000, Pakistan

ARTICLE INFO

Keywords:
Air quality parameters
Random forest
Machine learning-based estimation
Aspen hysys
Support vector machine
Environmental management operations

ABSTRACT

The prediction of relative humidity is a challenging task because of its nonlinear nature. The machine learning-based prediction strategies have attained significant attention in tackling a broad class of challenging nonlinear and complex problems. The random forest algorithm is a well-proven machine learning algorithm due to its ease of training and implementation, as it requires minimal preprocessing. The random forest algorithm has hitherto not been employed for estimating air quality parameters, such as relative humidity. In this study, the random forest approach is implemented to estimate the relative humidity as a function of dry- and wet-bulb temperatures. A well-known commercial process simulator called Aspen HYSYS® V10 is linked with MATLAB® version 2019a to establish a data mining environment. The robustness of the prediction model is evaluated against varying wet-bulb depressions. There is high absolute deviation that indicates a lower prediction performance of the model against the higher wet-bulb depression i.e., ~20.0 °C. The random forest model can predict relative humidity with a 1.1% mean absolute deviation compared to the values obtained through Aspen HYSYS. The performance of the RF estimation model is also compared with a well-known support vector regression model. The random forest model demonstrates 74.4% better performance than the support vector machine model for the problem of interest, i.e., relative humidity estimation. This study will significantly help the practitioners in efficient designing of air-dependent energy systems as well as in better environmental management through rigorous prediction of relative humidity.

1. Introduction

Rigorous estimation of air quality parameters, such as absolute humidity, relative humidity (RH), dry-bulb temperature (DBT), and wet-bulb temperature (WBT), is indispensable to maintain the smooth operation and performance of air-dependent energy systems. These systems are significantly affected by the fluctuations in air-quality parameters, particularly the RH. The variation in RH has thermodynamic effects on the performance of natural gas liquefaction processes employing air-based inter-stage coolers (Qyyum et al., 2018). The performance of photovoltaic solar module is also affected by the ambient absolute and RH (Sohani et al., 2020). The fracture properties of the rubberized concrete have been investigated under different humidity and temperature conditions (Wang et al., 2020). The performance of oscillating water column is also affected by the humidity (Medina-Lopez et al., 2019). Cooling towers are essential for thermal power plants and the petrochemical industry (Song et al., 2020). These towers are used to remove heat from processed water by evaporative cooling through ambient air. Air dryer using waste heat of the heating, ventilation, and air-conditioning (HVAC) systems is another benefit of rigorous prediction of air quality parameters (Ramadan et al., 2019).

Without accurate knowledge and estimation of the psychrometric parameters, HVAC engineers cannot design and select appropriate equipment with the optimal operating parameters for a system of interest. Hence, rigorous and robust estimation of RH is crucial for various aspects of daily life and industrial applications. Several investigations have been carried out to predict RH while considering various applications and reference data. For instance (Upreti and Ojha, 2017), used the Lawrence analytical model (Lawrence, 2005) for estimating the RHs for

* Corresponding author.
** Corresponding author.
E-mail addresses: maqyyum@yu.ac.kr (M.A. Qyyum), mynlee@yu.ac.kr (M. Lee).

https://doi.org/10.1016/j.jenvman.2021.112736
Received 28 January 2021; Received in revised form 18 April 2021; Accepted 29 April 2021
Available online 13 May 2021
0301-4797/© 2021 Elsevier Ltd. All rights reserved.

Nomenclature and abbreviations

HVAC	Heating, ventilation, and air conditioning
DBT	Dry-bulb temperature
WBT	Wet-bulb temperature
RH	Relative humidity
WBD	Wet-bulb depression
RF	Random forest
Ntree	Number of trees in forest
Nleaves	Number of leaves in a tree
Nsplits	Number of splits in a tree
Nfeature	Number of random features at each split
OOB	Out-of-bag data
ID3	Iterative Dichotomiser 3 algorithm
SVM	Support vector machine
RH_{Hysys}	RH calculated from Aspen Hysys
RH_{RF}	RH calculated from RF model
RH_{SVM}	RH calculated from SVM

hydrological processes such as evaporation and evapotranspiration (Bahadori et al., 2013). used WBD and DBT as independent variables to predict the RH at standard atmospheric conditions using a Vandermonde matrix predictive tool.

The random forest (RF) method is a well-proven and fully established machine learning approach (Ferreira et al., 2021; Pourghasemi et al., 2020; Reis et al., 2018). For training data selection, RF does not need any statistical pre-assumption, and it gives a more reliable prediction through the average results of all decision trees (Ghosh and Das, 2020). Many complex and nonlinear estimation problems have been solved successfully by using the RF approach. For instance (Ma and Cheng, 2016), used the RF model to identify the impacts of different possible features on the regional energy usage intensity of residential buildings (Becker and Thrän, 2017). used the RF and k-nearest neighbors' approach to complete wind turbine data to facilitate the wind integration studies (Li et al., 2018). used the RF regression model for online capacity estimation of lithium-ion batteries (Smarra et al., 2018). presented a RF-based data-driven predictive control model for climate control and building energy optimization (Prasad et al., 2019). designed a multi-stage machine learning model using the RF approach in coupling with the ant colony optimization algorithm to predict the monthly solar radiations. Several other studies have also solved complex estimation problems related to solar radiation prediction for tilted as well as horizontal surfaces using RF approach (Assouline et al., 2018; Hassan et al., 2017; Lou et al., 2016; Ramli et al., 2015). The RF algorithm has also been used to model and estimate the short-term air pollution effects in Wrocław based on meteorological and traffic conditions (Kamińska, 2018).

Machine learning-based models have been extensively employed for environmental data modeling and prediction (Ağbulut et al., 2020; Assouline et al., 2018; Li et al., 2020; Yang et al., 2020). However, until now, estimation problems relevant to air quality parameters (such as RH, the WBT, and the dew point temperature) have not been solved through any well-known machine learning algorithms, such as the RF.

2. Problem statement: relative humidity estimation

Investigations related to RH estimation are limited owing to the highly complex and nonlinear nature of RH estimation. The RH trend when the DBT and WBT are varied simultaneously is shown in Fig. 1. Accordingly, a curved surface-like pattern is observed, showing a highly nonlinear and complex RH relationship with the DBT and WBT.

Fig. 1(a) and (b) show the tilted and front sides of the spread scattered data of RH that may indicate that RH estimation follows a complicated (inexplicit) polynomial function. Therefore, to solve this complex estimation problem, new sophisticated approaches, such as machine learning-based prediction models, need to be explored.

To the best of the authors' knowledge, RH estimation, especially about the energy and process industries, has not yet been considered in open literature. This could be attributed to RH data scarcity corresponding to the DBT and WBT in open literature. Further, this leads to RH estimation turning into a difficult task for any process system of interest and environmental management-related applications. Therefore, this study serves to be the first one to consider the process

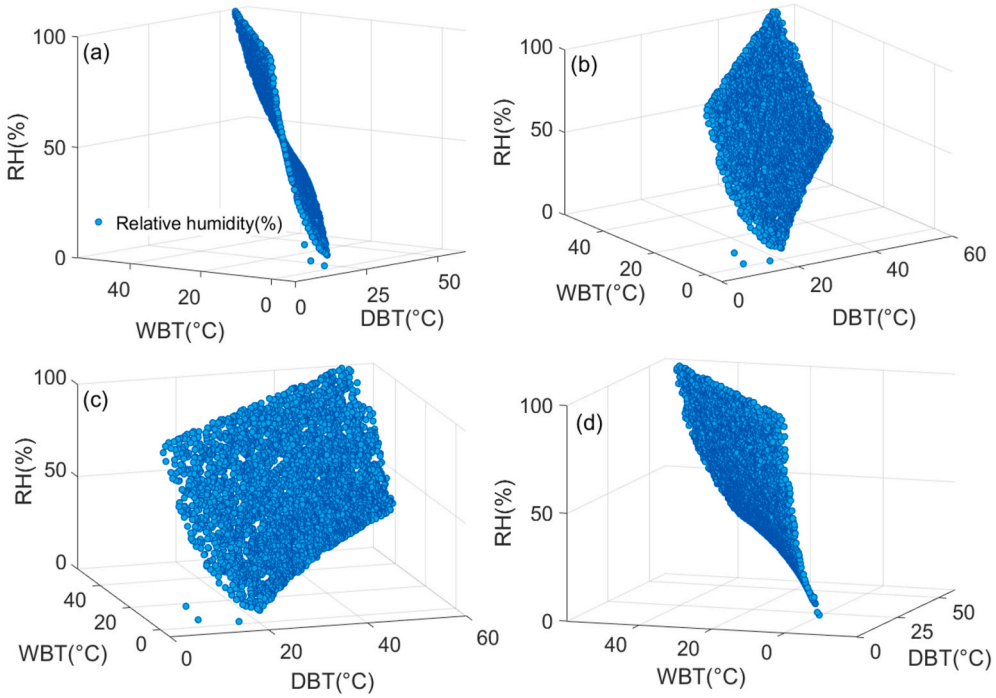

Fig. 1. Relative humidity trend against varying dry-bulb and wet-bulb temperatures.

engineering perspective for RH estimation. The major contributions of the proposed study include the following:

- A well-known process simulator called Aspen HYSYS® V10 is used to extract data for RH estimation as a function of the DBT and WBT.
- For data extraction, Aspen HYSYS® V10 is linked with MATLAB® version 2019a through the ActiveX server functionality.
- An RF-based estimation tool is established for RH estimation.
- The determination of the importance of the out-of-bag (OOB) features along with outlier measurement tests are performed on RH estimation training datasets.
- The robustness of the RF model for RH estimation is examined by varying the WBD.
- The RH estimation problem is also solved using a support vector machine (SVM) to evaluate the performance of the RF prediction model in comparison with SVM.

3. Methodology

The conceptual framework of the proposed study is schematized in Fig. 2.

3.1. Data mining environment

To address the issues associated with data collection for RH estimation, especially for process systems engineering applications, the Aspen HYSYS® V10 process simulator was used to extract the proposed study data. A humidification environment was simulated using a unit

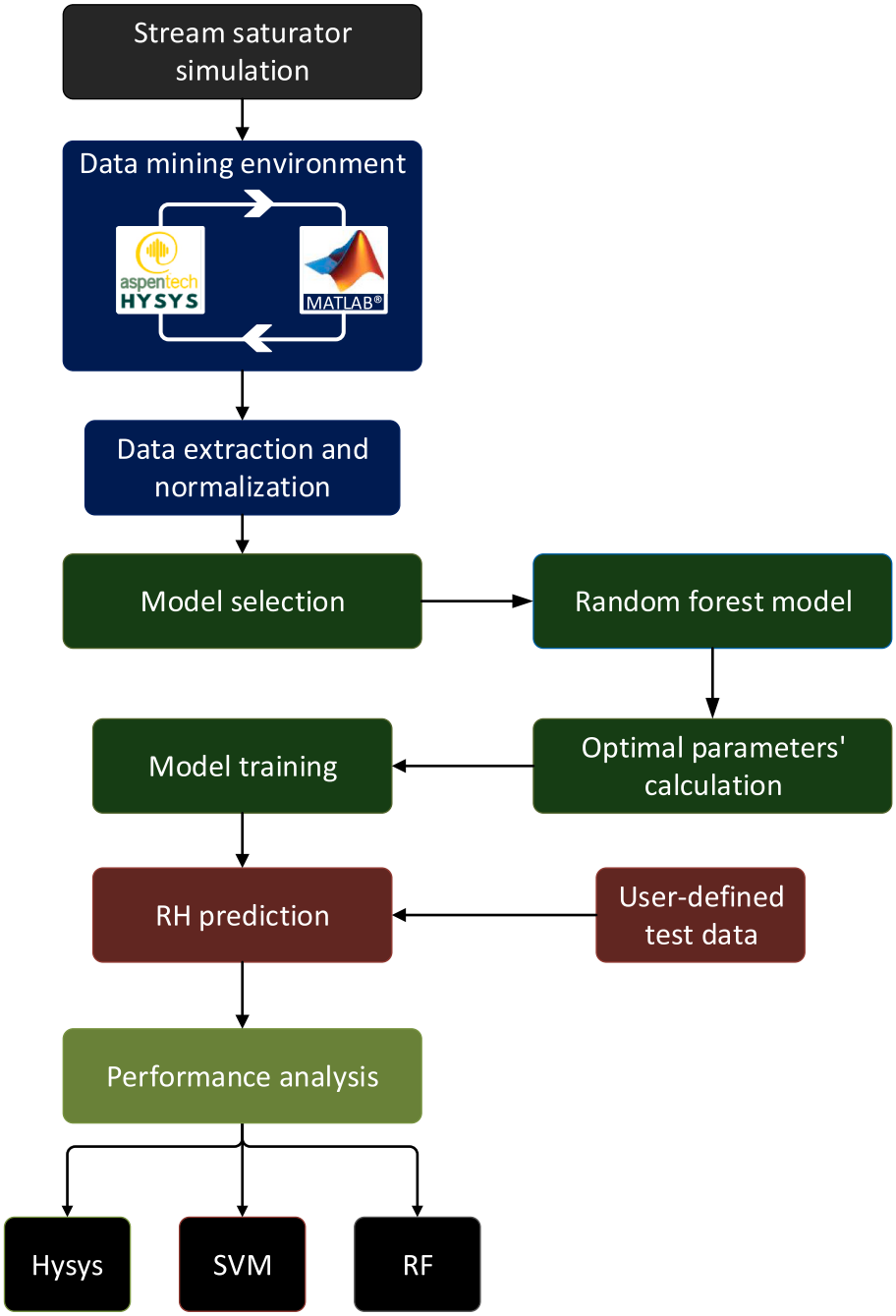

Fig. 2. Conceptual framework of the relative humidity estimation using a machine learning approach called the random forest.

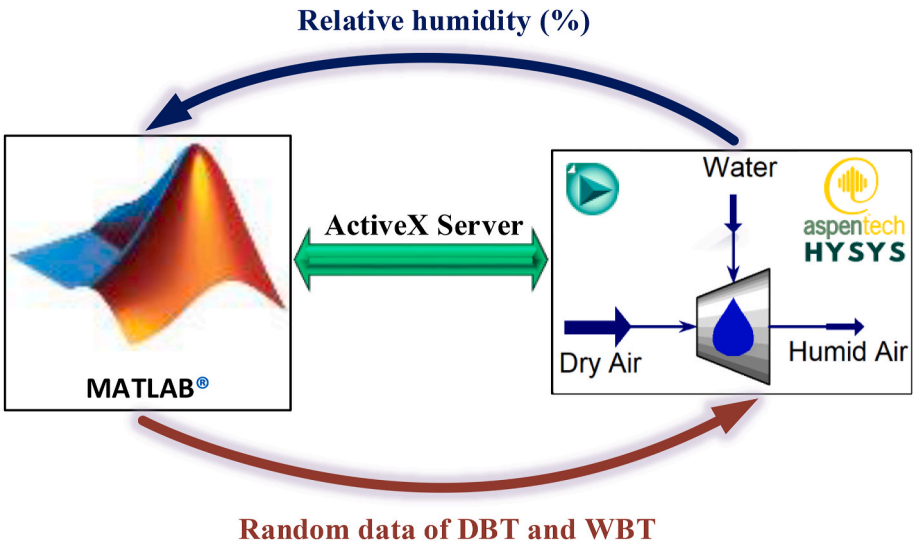

Fig. 3. Aspen HYSYS–MATLAB-based data mining environment.

Table 1
Conditions and assumptions for the development of data mining environment.

Property	Value
Pressure	1.0 atm
Dry air molar flow	100.0 kmole/h
Thermodynamic package	Antoine
Constraints for data extraction	
DBT	1 °C–60 °C
WBT	0 °C–55 °C
WBD	1–20.0 °C
RH	0–100%
DBT < WBT	Should not follow

operation called a "stream saturator" from the model palette of Aspen HYSYS. The RH data corresponding to the DBT and WBT were collected by linking the simulator with MATLAB® version 2019a, as shown in Fig. 3. Table 1 lists the primary conditions and assumptions that were used to build the data mining environment.

3.2. Random forest regression model

The prediction model for RH estimation was developed using a well-established RF algorithm with proven potential for classification and regression. The major features of the RF algorithm are its ability to perform robust predictions with no repetitive training, exploit all

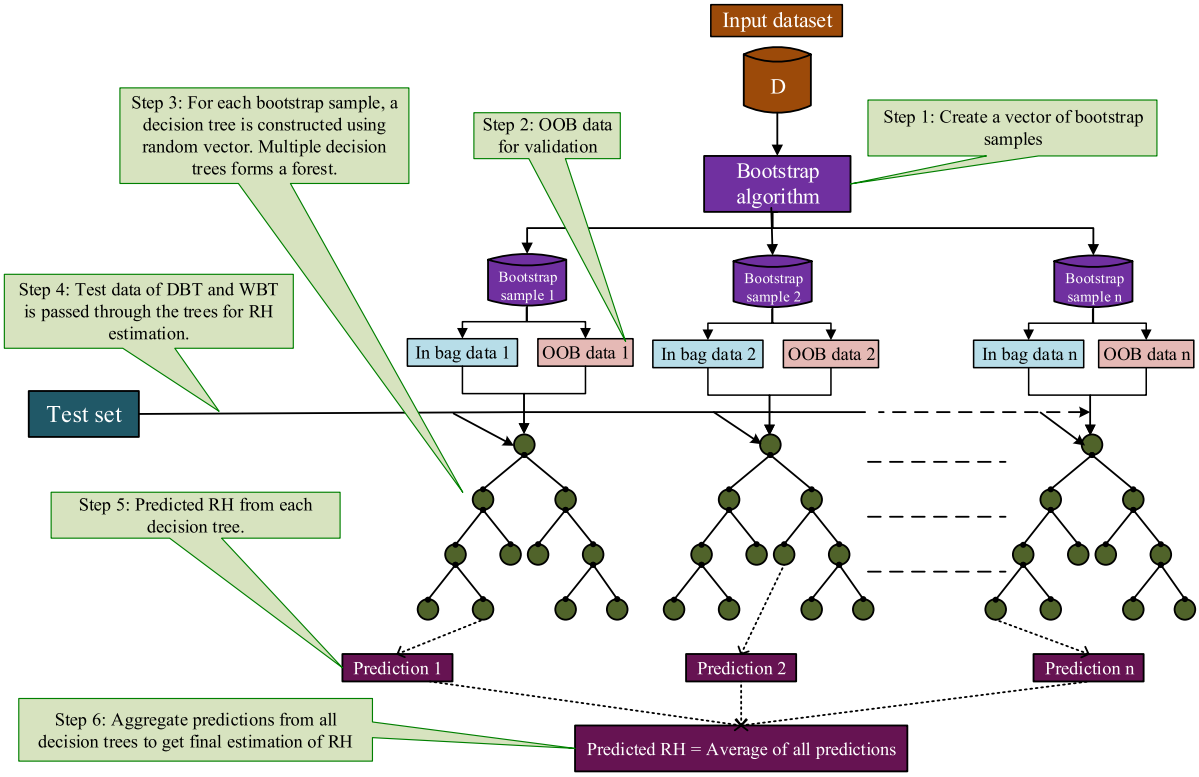

Fig. 4. Steps involved in building RF model for relative humidity estimation.

Table 2
Summery of steps involved in building RF model.

Step 1: Draw *n bootstrap* samples from the input dataset.
Step 2: About 1/3 elements of the *bootstrap* sample are considered as OOB data for the validation of tree's performance.
Step 3: An unpruned decision tree is constructed on about 2/3 data elements (In-bag data) of the *bootstrap* subset.
Step 4: Test data is passed through each tree and the prediction is made at leaf node.
Step 5: Prediction of each decision tree is obtained
Step 6: Final prediction on test data is obtained by aggregating the predictions of *n* trees.

possible combinations of input data sets, and handle unbalanced and high dimensional data while maintaining a stable performance (Archer and Kimes, 2008; Gokgoz and Subasi, 2015; Kursa, 2014; Statnikov et al., 2008).

The basic working structure of the RF model for RH estimation is shown in Fig. 4. Table 2 lists the major steps involved in the RF model adaptation for any particular estimation.

The proposed method of predicting RH as a function of the DBT and WBT is a multiple regression problem. Therefore, the decision trees in the RF model are treated as regression functions. A decision tree, first introduced by (Breiman et al., 1984) is built during the training process and depends on the complexity of the training data set. The core algorithm used for building decision trees is the Iterative Dichotomiser 3 (ID3) (Quinlan, 1986). The final output of the RF model is the average of the outputs of all the decision trees. A RF is an ensemble learning method for classification or regression where all the generated decision trees are combined using the bagging algorithm. Bagging, introduced by (Breiman, 1996), can improve the prediction performance through variance reduction. The RF model is built by randomly drawing subsamples from the input data set, building decision trees for them, and, finally, combining all the decision trees. The process of randomly collecting a subsample is known as "*bootstrap*." Thus, a *bootstrap* sample is obtained by randomly selecting *n* instances with replacement from the input data set. Approximately one-third of the sample is left out during construction, which is called the OOB data. Each time a regression tree is built, the OOB data is used to evaluate the regression tree's performance. The OOB data provides a running unbiased estimate of the prediction error as trees are added to the forest in the construction phase (Breiman, 2001). The OOB feature behaves like a built-in validation method of the RF algorithm. The total learning error can be obtained by averaging the prediction error of each tree.

The bagging algorithm is employed to generate several *bootstrap* samples to construct the respective prediction trees. The ensemble produces the output corresponding to each tree. Consequently, the final estimation is obtained through an average aggregation of the predictions of all the trees (Rodriguez-Galiano et al., 2015). The major advantage of bagging is the increase in the diversity of trees because they are built from different data subsets that in turn, further reduces the probability of correlation of trees (Rodriguez-Galiano et al., 2015). Some data may be used multiple times for tree construction, while some may never be used at all. Thus, bagging increases the stability of the RF. It also makes the RF algorithm more robust to slight variations in the input data set (Breiman, 2001). Another advantage of using bagging is its immunity to noisy predictors. A weak predictor is more susceptible to noise; however, due to the consideration of average outputs from several un-correlated decision trees, the noise sensitivity can decrease (Lahouar and Ben Hadj Slama, 2017). Furthermore, the growth of trees without pruning is another vital feature of the RF that reduces computational burdens (Rodriguez-Galiano et al., 2015).

3.3. Implementation of random forest model

In this study, the RF-based model was implemented using MATLAB® version 2019a to solve the RH estimation problem. RH was considered a "response variable", whereas the DBT and WBT were considered "predictor variables." A *TreeBagger* ensemble feature was employed for training and testing the RF model.

To solve the proposed estimation problem, the optimal parameter values involved in the RF model need to be determined and fixed. The major parameters of the RF model are the number of trees (*Ntree*), number of leaves in a tree (*Nleaves*), number of splits in a tree (*Nsplits*), and number of random features (*Nfeature*) at each split. The optimal

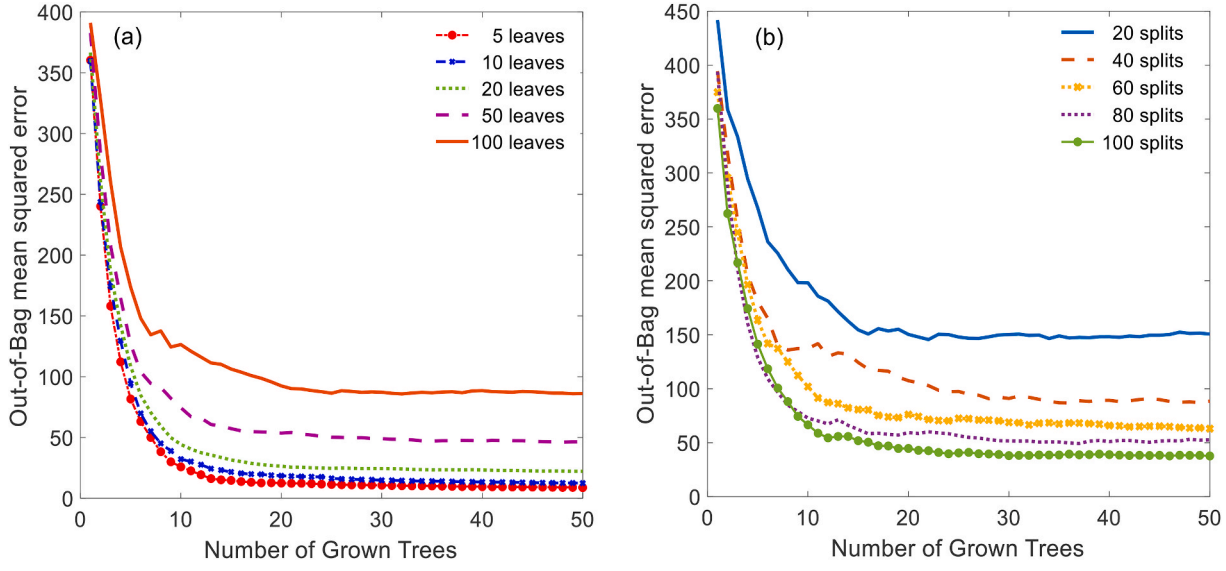

Fig. 5. Sensitivity analysis for the optimal selection of the RF model parameters; (a) number of leaves in a tree; (b) number of splits in a tree.

values of these parameters can be determined either by performing sensitivity analyses on the data set under examination (Liaw and Wiener, 2001) or through the application of any optimization algorithm, such as a genetic algorithm (GA), particle swarm optimization (PSO) algorithm, or vortex search optimization (VSO) algorithm. However, in the proposed study, a conventional sensitivity analysis was performed to determine the optimal values of the RF model parameters. Fig. 5 serves as a guide for choosing the optimal values of the RF model parameters. Irrespective of the size of *Nleaves* and *Nsplits*, the OOB mean squared error is meaningless when the number of trees in the RF is over 200. It can also be concluded from Fig. 5(a) and (b) that the OOB error is minimum when *Nleaves* = 5 and *Nsplits* = 100, respectively.

Nfeatures is a sensitive parameter that determines the strength of a tree and its correlation with other trees. When *Nfeatures* is increased, the strength of a tree increases, along with a simultaneous increase in the trees' correlation. Further, although the prediction performance of the model is improved by increasing the strength of the tree, it weakens when the correlation between the trees increases. It has been reported (Liaw and Wiener, 2001) that maintaining *Nfeatures* as one-third of the total number of predictor variables would be appropriate. For the RH prediction problem, as there are only two predictor variables, the value of *Nfeatures* was selected as 1.0. Furthermore, Table 3 lists the optimum values of the RF model parameters for solving the RH estimation problem.

Table 3
Optimum parameters of RF model for RH estimation.

Parameter	Description	Value
Ntrees	Total number of trees grown in forest	200.0
Nleaves	Total number of leaves in tree	5.0
Nsplits	Total number of splits in tree	100.0
Nfeatures	Number of random features at each split	1.0

The number of trees grown in a forest is directly proportional to the accuracy and robustness of prediction. However, increasing the number of trees in the forest can lead to an excessive computational burden. The generalization error converges with an increase in the number of trees. This implies that after reaching a specific number of trees, the prediction performance remains unaffected. The value of the number of trees is maintained sufficiently high to ensure error convergence.

The RF model would be trained based on the training data by fixing the optimum parameters, as listed in Table 3. For this, a data set of 3500 observations was obtained by establishing a data mining environment described in Section 2. A forest (*TreeBagger*) was created on the training data set by constructing 200 *bootstrap* samples with replacement. A decision tree against each *bootstrap* sample was constructed with *Nleaves* and *Nsplits* values of 5 and 100, respectively. One of the regression trees is shown in Fig. 6 (zoomed part of Fig. 4). Thus, the RF is ready to make predictions on new (user-defined) data. Therefore, the testing data is distributed in the forest to begin the prediction procedure. The data flows into the decision trees and is traced to the constructed splits to output the respective predictions. The final prediction is obtained by determining the average aggregation of the predictions for all the trees.

4. Results and discussion

4.1. Diversity analysis of training data

The prediction performance of a learning-based algorithm strongly depends on the training data quality. Generally, the outlier measurement test is used to analyze the diversity in the training data set observations. It identifies and measures the number of strange observations in the training data. It is essential to evaluate the data obtained from the process simulator Aspen HYSYS® for the proposed RH estimation. In this study, the outlier measurement test was performed on the data set of 3500 observations, as shown in Fig. 7. The used data looks appreciable because of its versatile nature considering the fundamental

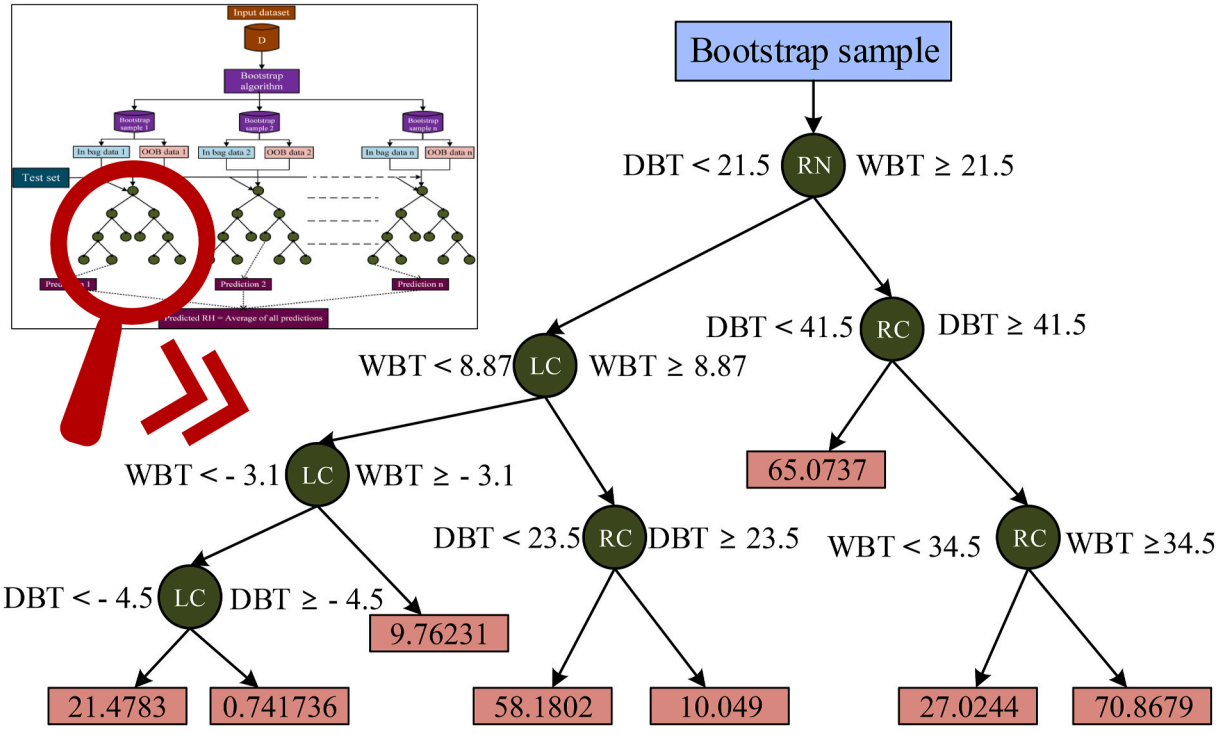

Fig. 6. Sample of regression tree in building RF model for RH estimation.

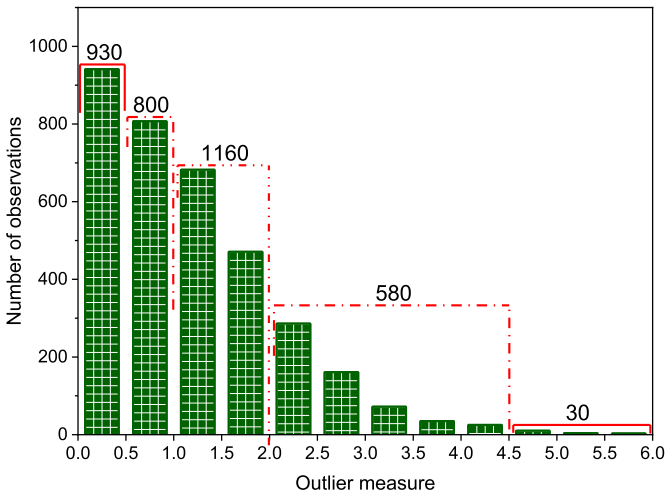

Fig. 7. Results of outlier measurement test on training dataset for RH estimation.

constraints (see Table 1) of the DBT and WBT.

To observe the diversity in the used data, the number of observations is classified according to the different values of the outlier measurements, such as 0–0.5, 0.5–1.0, 1.0–2.0, 2.0–4.5, and 4.5–6.0. It can be observed that the outlier value of 26.57% (930 out of 3500) observations is between 0–0.5. Similarly, 800, 1160, and 580 observations of the training data were found within outlier values of 0.5–1.0, 1.0–2.0, and 2.0–4.5, respectively. Training data with a higher outlier measurement (i.e., 4.5–6.0) was also observed, which comprised approximately 30 samples. These lower to higher outlier observations are responsible for the diversity of the used training data set.

4.2. Out-of-bag features analysis

The performance of the RF model for RH estimation was observed based on the different training data sets, namely, 1500, 2500, and 3500 datasets. However, before evaluating the RF model, the importance of OOB feature for the three categories of training data sets (1500, 2500, and 3500) was determined, as shown in Fig. 8.

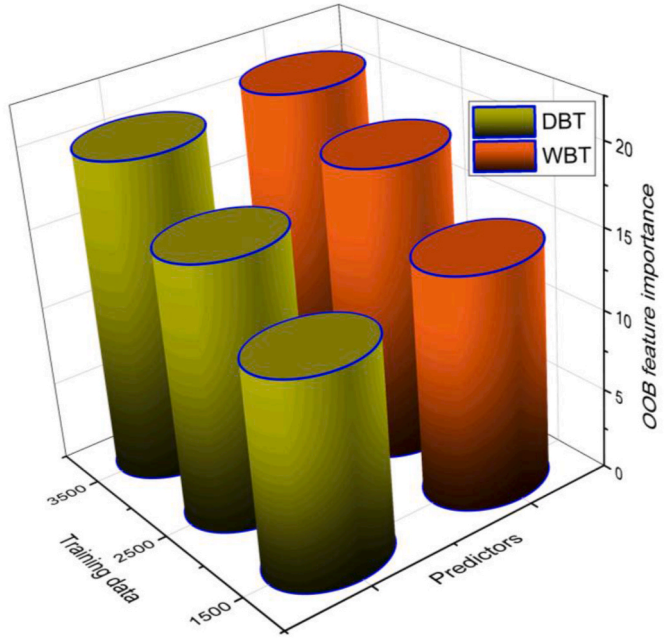

Fig. 8. OOB feature importance of dry-bulb and wet-bulb temperature under different training datasets.

Although both DBT and WBT are mandatory requirements for the prediction of RH, WBT has a comparatively higher OOB feature importance when compared to DBT. However, it is noteworthy that the difference between the OOB feature importance values of DBT and WBT decreased upon increasing the no. of observations in a particular training data. Therefore, it is hard to conclude as to which predictor impacts the estimation of RH the most. Fig. 8 also indicates the possibility of the OOB feature importance values of DBT and WBT becoming equal when the training data set size exceeds 3500 observations.

4.3. Performance analysis

The RF model was trained using a training data of 1500 observations and evaluated compared to the RH values obtained from HYSYS. Similarly, it was trained using the data sets of 2500 and 3500 observations as well. Considering the process engineering perspective, the Aspen HYSYS® V10 version was chosen as a benchmark to evaluate the prediction performance of the RF model for RH estimation. A total of 15 random user-defined values of the predictor variables (DBT and WBT) were considered for evaluating the RF model's performance. These user-defined values were not included in the observations of the primary training data set. Table 4 lists the estimated RH values for user defined DBT and WBT values under varying training data sets in comparison with Aspen HYSYS. Table 4 demonstrates that the prediction performance of the RF model is enhanced upon increasing the number of training data sets. However, it has been reported that the size of the training data set should be at least 3000 to 5000 to attain rigorous and robust prediction. Therefore, in this study, the training data set used had a size of up to 3500 observations. The mean absolute deviation was 1.1% for a training data set of 3500 observations, whereas it was 2.4% and 1.6% for the training data sets with 1500 and 2500 observations, respectively. It is noteworthy that there was a 33% performance improvement when the size of the training data set was increased from 1500 to 2500. Moreover, when the observations in a training data set size increased from 2500 to 3500, the prediction performance was further improved by 31.2%, which is 1.8% lower than the performance improvement from 1500 to 2500. In summary, an overall performance improvement of 54.2% was observed when the training observations were increased from 1500 to 3500. It can be concluded that the prediction performance of the RF model can be further improved by increasing the number of observations in the training data set.

Table 4
RH estimated through RF model at user-defined DBT and WBT values under varying observations in a training data set.

DBT (°C)	WBT (°C)	RH_{Hysys} (%)	1500 observations		2500 observations		3500 observations	
			RH_{RF} (%)	ε (%)	RH_{RF} (%)	ε (%)	RH_{RF} (%)	ε (%)
55.5	43.9	50.4	51.3	2.1	51.7	2.7	51.2	1.61
31.4	22.2	52.0	51.7	0.6	52.6	1.2	52.5	0.94
23.1	17.8	67.0	68.8	2.6	68.5	2.1	67.7	1.06
9.5	7.2	80.8	79.2	1.9	79.6	1.4	79.8	1.24
39.4	32.8	65.7	65.2	0.8	65.4	0.5	65.2	0.80
30.8	16.4	33.0	34.3	3.8	33.7	2.2	33.4	1.24
39.2	21.3	28.1	30.3	7.9	28.8	2.7	28.6	1.97
29.3	23.5	66.7	69.0	3.3	66.7	0.0	66.7	0.08
13.9	9.3	65.9	64.1	2.7	67.9	3.0	67.3	2.11
41.4	30.1	48.3	46.4	4.1	46.8	3.1	47.8	1.12
24.1	21.1	80.7	80.7	0.0	80.4	0.4	80.5	0.30
42.3	41.1	92.8	90.0	2.9	90.7	2.3	91.2	1.66
48.5	42.0	67.7	66.8	1.2	67.0	0.9	67.1	0.86
41.3	35.3	68.8	69.5	1.0	69.6	1.2	69.1	0.42
39.6	36.3	81.7	82.5	1.1	81.1	0.6	80.8	1.09
Mean absolute deviation percent				**2.4**		**1.6**		**1.1**

$\varepsilon = |((RH_{Hysys}–RH_{RF})/RH_{Hysys})*100|$.

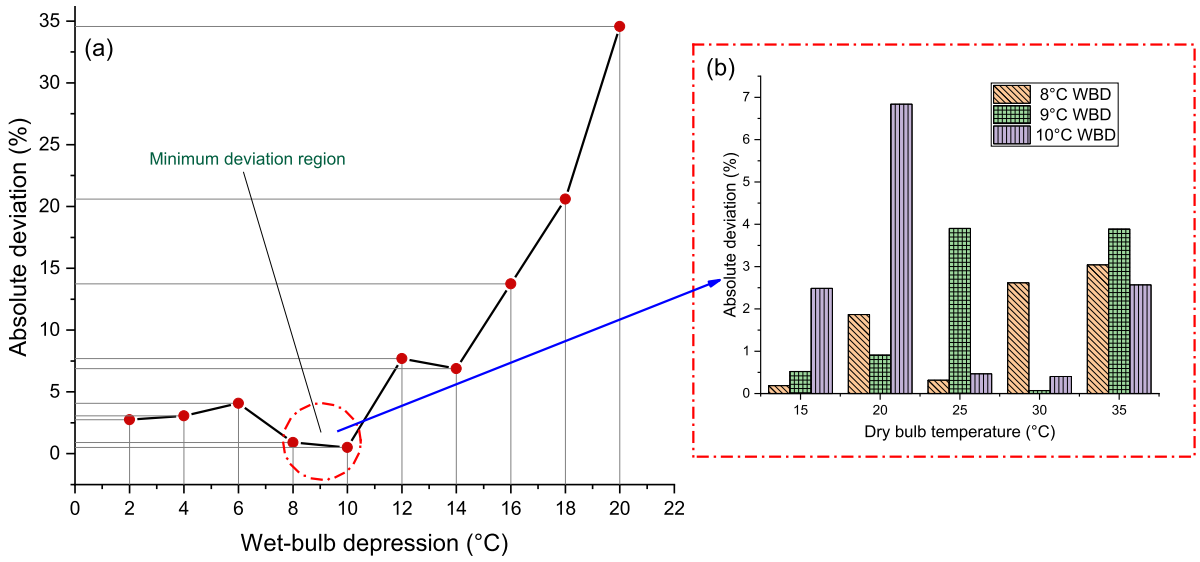

Fig. 9. Robustness evaluation of random forest model for relative humidity estimation on used training dataset.

The robustness of the proposed RH estimation model was also examined and analyzed. The absolute deviation (%) of the estimated RH was compared to HYSYS-RH under varying WBD (difference of DBT and WBT) conditions. Fig. 9(a) presents the absolute deviation in the estimated RH corresponding to WBD. The absolute deviation percent shows an overall increasing trend when the WBD increases. The minimum absolute deviation was observed when the WBD value was 8.0–10.0 °C, as shown by the red-circled region in Fig. 9(a). This minimum deviation region was further analyzed by varying the DBT, as shown in Fig. 9(b). As the WBD increased, the RH estimation problem turned more complicated and nonlinear in nature, which ultimately led to the reduction in the prediction performance of the RF model. Further, higher values of the WBD make it difficult for the RF model to find correlation between the variables at the given training samples of 3500.

Prior to the red-circled region in Fig. 9(a) being analyzed, it was assumed that the proposed RH estimation model predicted accurate RH values whenever the WBD was 8.0–10.0 °C. However, the analysis shown in Fig. 9(b) revealed that it is unnecessary for the absolute deviation to always be $\leq$ 2.0% for any DBT when the WBD value is 8.0–10.0 °C. These analysis results highlight the importance of data and its quality to take maximum potential benefits of any learning-based model. The prediction performance of the RF model can be improved by increasing the accurate training observations (or samples) especially for higher WBD. However, the addition of any unrealistic observation in the training data set may have a reverse effect on the performance.

The prediction performance of the proposed RF model was also compared with performing a well-known support vector machine (SVM) regression model. For a fair comparison, the SVM regression model was also trained with the same training data set of 3500 observations. The prediction performance comparison of the RF model and SVM model for RH estimation is listed in Table 5.

It was observed that the overall performance of the RF model was 74.4% higher than that of the SVM model. However, there were few interesting values (given in Table 5) of DBT i.e., 23.1°C, 30.8°C, 13.9°C, and 39.6°C corresponding to WBT values of 17.8°C, 16.4°C, 9.3°C, and 36.3°C, respectively where the prediction performance of the SVM model was higher than that of the RF model. It cannot be concluded that the RF model will always perform better than the SVM regression model; the performance quality might reverse depending on the specific problem and quality of the training samples. The performance of almost all machine learning models mainly depends on the following factors:

Table 5
Performance analysis comparison of the RF prediction with SVM model.

DBT (°C)	WBT (°C)	RH_{Hysys} (%)	RH_{SVM} (%)	RH_{RF} (%)	ε^* (%)	ε^{**} (%)
55.5	43.9	50.4	53.9	51.2	7.0	1.61
31.4	22.2	52.0	54.1	52.5	4.1	0.94
23.1	17.8	67.0	67.0	67.7	0.0	1.06
9.5	7.2	80.8	74.0	79.8	8.4	1.24
39.4	32.8	65.7	67.7	65.2	3.0	0.80
30.8	16.4	33.0	33.2	33.4	0.7	1.24
39.2	21.3	28.1	22.2	28.6	20.7	1.97
29.3	23.5	66.7	67.3	66.7	0.8	0.08
13.9	9.3	65.9	66.0	67.3	0.2	2.11
41.4	30.1	48.3	49.7	47.8	2.9	1.12
24.1	21.1	80.7	76.8	80.5	4.9	0.30
42.3	41.1	92.8	90.4	91.2	2.5	1.66
48.5	42.0	67.7	71.5	67.1	5.6	0.86
41.3	35.3	68.8	70.9	69.1	3.0	0.42
39.6	36.3	81.7	81.3	80.8	0.5	1.09
Mean absolute deviation (%)					**4.3**	**1.1**

$\varepsilon^* = (|(RH_{Hysys} - RH_{SVM})|/RH_{Hysys})*100$, $\varepsilon^{**} = (|(RH_{Hysys} - RH_{RF})|/RH_{Hysys})*100$.

- The size of training data
- Quality (in terms of correctness) of training data
- Problem complexity
- Interdependence of predictors
- The nature of correlation among the variables
- Tunning of model's parameters

Some algorithms work well on less training dataset (e.g., SVM) and some on higher training dataset (e.g., RF). In our study, RF model was evaluated on maximum 3500 training samples. The prediction performance of the RF may be improved by adding more and accurate training observations. Addition of any unrealistic observation in the training data set may reversely affect the performance. In summary, if we apply more than one algorithm on same problem, the prediction performance may vary by varying above factors.

5. Practical implications of this study

Generally, the behavior of RH corresponding to any DBT and WBT with a high or low WBD is uncertain. For industrial applications (as mentioned in the introduction section), air with a higher DBT (e.g., reheating air for drying purposes) can be used, thereby leading to higher

values of WBD. However, analyzing the air quality parameters for industrial applications rather than environmental forecasting remains an open issue. Therefore, still, the RH does not consider rigorously especially in designing stage of low/high pressure air-based energy systems or heat exchangers. The major reason is the unavailability of data at different operating pressures other than atmospheric. Although, the RF regression model has been successfully implemented to estimate the RH corresponding to a wide range of DBT and WBD, but practically its performance totally depends on the quality and quantity of training data. For process systems engineering applications, the training data can be obtained through any commercial simulators such as Aspen Hysys, Aspen Plus, Pro-II, and Unisim etc. Nevertheless, the reliable and big data extraction for any air-based energy system is not an easy exercise, mainly due to the highly nonlinear thermodynamic interactions involved.

In this study, the performance of the RF estimation model has been examined using different training data sets comprising 1500, 2500, and 3500 observations. The OOB feature importance and outlier measurement tests were performed to analyze the quality of the training dataset. Finally, the minimum absolute deviation in the predicted RH was determined at the training data set of 3500 samples. The robustness of the RF model under varying WBD has also been evaluated. The prediction performance of the RF model can be improved by increasing the accurate training observations (or samples) especially for higher WBD, as observed from the presented performance analysis (see Table 4). However, the addition of any unrealistic observation in the training data set may have a reverse effect on the performance.

This study helps process engineers designing different air-dependent industrial operations such as drying, cooling, air gasification, and heating. However, from a process system engineering perspective, several issues still exist related to RH estimation, such as the DBT and WBT measurements and calculations in freezing ($<$0.0 °C) and heating ($>$50.0 °C) conditions. It is essential to be aware of reliable data and calculation formulas for determining vapor pressures at extreme conditions. The well-known Antoine model might also need to be modified for it to extend its application range. Other thermodynamic models (e.g., Peng-Robinson) that could tackle high pressure non-linear systems may also be investigated to extract the data for RH estimation using machine learning approaches in a wide-range of air-based industrial applications.

There are various approaches and instruments used in measuring ambient air quality parameters and amounts of pollutants. However, each approach and instrument gives a different value and has uncertain performance, mainly due to the complex nature of environmental problems and pollutants' diversity. The apparatus should be uncovered to the environmental impact to advance the air pollutant measuring devices, such as variations in temperature, dustiness, vibrations, and humidity. Moreover, all air pollution technologies' performance strongly depends on the nature of the dust, the ambient temperature, and the humidity. Another important implication of this study is in efficient management of environment-related processes and operations. For instance, the balance between global evaporation and precipitation can be evaluated and maintained through rigorous estimation of air quality parameters, particularly RH. As reported in (Fath et al., 2020), air pollutants' registered values have to be correlated with the actual meteorological conditions such as ambient temperature and humidity. Therefore, this study could help to calibrate the instruments used for observing and determining the air pollutants.

6. Conclusions and future prospects

The RF model has sufficient potential to solve highly nonlinear estimation problems such as that of RH estimation. The prediction performance of the RF model is not impressive for few reasonable DBT and WBT values (e.g., at DBT = 13.9 °C and WBT = 9.3 °C with 2.11% absolute deviation), and the exact reason behind this relatively low performance could not be determined. If the RF model could be interpreted, this would be the biggest limitation associated with the RF model. The prediction performance of the RF model strongly depends on both the quantity and quality of the training data. At higher values (~20.0 °C) of WBD, there is high absolute deviation that indicates a lower prediction performance of the RF model against the higher WBD. This low performance of the RF model can be improved through rigorous optimizations (using either evolutionary or metaheuristic algorithms) of the RF model's parameters. Based on the comparison of the RF model with SVM, a conclusion cannot be reached about the performance superiority of the RF model. The mean performance of the RF model is better than that of the SVM for the problem of interest.

In the future, this study would extend to explore the RF model for the relative RH prediction under varying pressure values for different industrial applications. Other machine learning approaches can also be examined in comparison with the RF algorithm using the same training dataset. Moreover, better performance of the RF algorithm can be achieved using some state-of-the-art deterministic optimizers to find the optimal values of the model's parameters. There is wide scope in implementing RF models for the prediction of design variables of various complex energy systems i.e., natural gas liquefaction process, hydrogen liquefaction process, and integrated energy systems.

Credit author statement

Kinza Qadeer: Conceptualization, methodology, investigation, and software. Ashfaq Ahmad: Methodology, investigation, data curation, and software. Muhammad Abdul Qyyum: Conceptualization, Writing - Original Draft and software, and Supervision. Abdul-Sattar Nizami: Writing - Review & Editing. Moonyong Lee: Supervision, Writing - Review & Editing, and 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.

Acknowledgment

This work was supported by the Priority Research Centers Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (2014R1A6A1031189).

References

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 52, 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Assouline, D., Mohajeri, N., Scartezzini, J.-L., 2018. Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests. Appl. Energy 217, 189–211. https://doi.org/10.1016/J.APENERGY.2018.02.118.

Ağbulut, Ü., Gürel, A.E., Ergün, A., Ceylan, İ., 2020. Performance assessment of a V-trough photovoltaic system and prediction of power output with different machine learning algorithms. J. Clean. Prod. 268, 122269. https://doi.org/10.1016/j.jclepro.2020.122269.

Bahadori, A., Zahedi, G., Zendehboudi, S., Hooman, K., 2013. Simple predictive tool to estimate relative humidity using wet bulb depression and dry bulb temperature. Appl. Therm. Eng. 50, 511–515. https://doi.org/10.1016/j.applthermaleng.2012.07.033.

Becker, R., Thrän, D., 2017. Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors. Appl. Energy 208, 252–262. https://doi.org/10.1016/j.apenergy.2017.10.044.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/bf00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, Leo, Friedman, Jerome, Charles, J., Stone, R.A.O., 1984. Classification and Regression Trees. CRC Press.

Fath, B.D., Jørgensen, S.E., Cole, M., 2020. Managing Air Quality and Energy Systems, second ed. CRC Press, Boca Raton. https://doi.org/10.1201/9781003043461.

Ferreira, R.G., da Silva, D.D., Elesbon, A.A.A., Fernandes-Filho, E.I., Veloso, G.V., de Souza Fraga, M., Ferreira, L.B., 2021. Machine learning models for streamflow

regionalization in a tropical watershed. J. Environ. Manag. 280, 111713. https://doi.org/10.1016/j.jenvman.2020.111713.
Ghosh, S., Das, A., 2020. Wetland conversion risk assessment of East Kolkata Wetland: a Ramsar site using random forest and support vector machine model. J. Clean. Prod. 275, 123475. https://doi.org/10.1016/j.jclepro.2020.123475.
Gokgoz, E., Subasi, A., 2015. Comparison of decision tree algorithms for EMG signal classification using DWT. Biomed. Signal Process Contr. 18, 138–144. https://doi.org/10.1016/j.bspc.2014.12.005.
Hassan, M.A., Khalil, A., Kaseb, S., Kassem, M.A., 2017. Exploring the potential of tree-based ensemble methods in solar radiation modeling. Appl. Energy 203, 897–916. https://doi.org/10.1016/j.apenergy.2017.06.104.
Kamińska, J.A., 2018. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław. J. Environ. Manag. 217, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094.
Kursa, M.B., 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinf. 15 https://doi.org/10.1186/1471-2105-15-8.
Lahouar, A., Ben Hadj Slama, J., 2017. Hour-ahead wind power forecast based on random forests. Renew. Energy 109, 529–541. https://doi.org/10.1016/j.renene.2017.03.064.
Lawrence, M.G., 2005. The relationship between relative humidity and the dewpoint temperature in moist air: a simple conversion and applications. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/BAMS-86-2-225.
Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J.C.-W., van den Bossche, P., Mierlo, J. Van, Omar, N., 2018. Random forest regression for online capacity estimation of lithium-ion batteries. Appl. Energy 232, 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182.
Li, Z., Yim, S.H.-L., Ho, K.-F., 2020. High temporal resolution prediction of street-level PM2.5 and NOx concentrations using machine learning approach. J. Clean. Prod. 268, 121975. https://doi.org/10.1016/j.jclepro.2020.121975.
Liaw, A., Wiener, M., 2001. Classification and Regression by RandomForest.
Lou, S., Li, D.H.W., Lam, J.C., Chan, W.W.H., 2016. Prediction of diffuse solar irradiance using machine learning and multivariable regression. Appl. Energy 181, 367–374. https://doi.org/10.1016/j.apenergy.2016.08.093.
Ma, J., Cheng, J.C.P., 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Appl. Energy 183, 193–201. https://doi.org/10.1016/j.apenergy.2016.08.096.
Medina-Lopez, E., Borthwick, A.G.L., Moñino, A., 2019. Analytical and numerical simulations of an oscillating water column with humidity in the air chamber. J. Clean. Prod. 238, 117898. https://doi.org/10.1016/j.jclepro.2019.117898.
Pourghasemi, H.R., Sadhasivam, N., Yousefi, S., Tavangar, S., Ghaffari Nazarlou, H., Santosh, M., 2020. Using machine learning algorithms to map the groundwater recharge potential zones. J. Environ. Manag. 265, 110525. https://doi.org/10.1016/j.jenvman.2020.110525.
Prasad, R., Ali, M., Kwan, P., Khan, H., 2019. Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation. Appl. Energy 236, 778–792. https://doi.org/10.1016/j.apenergy.2018.12.034.
Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106. https://doi.org/10.1007/bf00116251.
Qyyum, M.A., Ali, W., Hussain, A., Bahadori, A., Lee, M., 2018. Feasibility study of environmental relative humidity through the thermodynamic effects on the performance of natural gas liquefaction process. Appl. Therm. Eng. 128, 51–63.
Ramadan, M., Murr, R., Khaled, M., Olabi, A.G., 2019. Air dryer using waste heat of HVAC systems – code development and experimental validation. Appl. Therm. Eng. 147, 302–311. https://doi.org/10.1016/j.applthermaleng.2018.10.087.
Ramli, M.A.M., Twaha, S., Al-Turki, Y.A., 2015. Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. Energy Convers. Manag. 105, 442–452. https://doi.org/10.1016/j.enconman.2015.07.083.
Reis, I., Baron, D., Shahaf, S., 2018. Probabilistic random forest: a machine learning algorithm for noisy data sets. Astron. J. 157, 16. https://doi.org/10.3847/1538-3881/aaf101.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.
Smarra, F., Jain, A., de Rubeis, T., Ambrosini, D., D'Innocenzo, A., Mangharam, R., 2018. Data-driven model predictive control using random forests for building energy optimization and climate control. Appl. Energy 226, 1252–1272. https://doi.org/10.1016/j.apenergy.2018.02.126.
Sohani, A., Shahverdian, M.H., Sayyaadi, H., Garcia, D.A., 2020. Impact of absolute and relative humidity on the performance of mono and poly crystalline silicon photovoltaics; applying artificial neural network. J. Clean. Prod. 276, 123016. https://doi.org/10.1016/j.jclepro.2020.123016.
Song, G., Zhi, X., Fan, F., Wang, W., Wang, P., 2020. Cooling performance of cylinder-frustum natural draft dry cooling tower. Appl. Therm. Eng. 180, 115797. https://doi.org/10.1016/j.applthermaleng.2020.115797.
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinf. https://doi.org/10.1186/1471-2105-9-319.
Upreti, H., Ojha, C.S.P., 2017. Estimation of relative humidity and dew Point temperature using Limited meteorological data. J. Irrigat. Drain. Eng. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001225.
Wang, J., Guo, Z., Zhang, P., Yuan, Q., Guan, Q., 2020. Fracture properties of rubberized concrete under different temperature and humidity conditions based on digital image correlation technique. J. Clean. Prod. 276, 124106. https://doi.org/10.1016/j.jclepro.2020.124106.
Yang, Q., Yuan, Q., Li, T., Yue, L., 2020. Mapping PM2.5 concentration at high resolution using a cascade random forest based downscaling model: evaluation and application. J. Clean. Prod. 277, 123887. https://doi.org/10.1016/j.jclepro.2020.123887.