Interpretable Soft Sensors using Extremely Randomized Trees and SHAP

Liang Cao [*] **Xiaolu Ji** [**] **Yankai Cao** [*] **Yupeng Li** [*,***]
Lim C. Siang [****] **Jin Li** [****] **Vijay Kumar Pediredla** [*]
R. Bhushan Gopaluni [*]

[*] *Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada (e-mail: bhushan.gopaluni@ubc.ca)*
[**] *Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*
[***] *School of Automation, China University of Geosciences, Wuhan, 430074, China*
[****] *Department of Process Control Engineering, Burnaby Refinery, Burnaby, BC, V5C 1L7, Canada*

Abstract: Tree-based ensemble models are easy to implement and have been widely used in various fields. However, they have limitations in industrial process applications since the majority of tree-based ensemble models are prone to over-fitting. In addition, the internal structure of tree-based ensemble models is very complex and the output of the model is also difficult to explain, which makes its application in industrial soft sensors very challenging. The purpose of this work is to build accurate and interpretable soft sensors for industrial processes. First, to deal with overfitting, a robust tree-based ensemble model and extremely randomized trees are used to build accurate soft sensors. Then, to improve model interpretability, an interpretable machine learning algorithm, namely Shapely additive explanation, is used to infer the global and local contributions of each feature to the predictions. Finally, the effectiveness of the proposed algorithms is validated on real industrial fluid catalytic cracker unit data.

Keywords: Interpretability, SHAP, Extremely Randomized Trees, Soft Sensor

1. INTRODUCTION

In modern industrial processes, it is necessary to monitor a large number of critical variables that are closely related to the process safety and economic benefits (Gopaluni et al. (2020)). These critical variables are called quality variables. However, some quality variables are difficult or expensive to measure by sensors in real-time, which poses challenges for the monitoring and control of industrial processes. To achieve real-time monitoring of quality variables, data-driven soft sensors are proposed to estimate the difficult or expensive to measure quality variables from easily measured process variables (Zhu et al. (2020)).

With the development of big data approaches and increasing computing power, tree-based models (including decision trees (DT) (Kotsiantis (2013)), random forests (RF) (Ho (1995)), light gradient boosting machine (Ke et al. (2017)), extremely randomized trees (ET) (Geurts et al. (2006)), etc.) have been widely used in various fields. Compared with statistical models, tree-based models are easy to implement, and the accuracy of the models is often significantly better. During the past decade, almost all winners of Netflix competitions, Kaggle competitions, etc., utilized tree-based ensemble models in their solutions.

However, these algorithms have limitations in industrial process applications since modern industrial processes are often characterized by high dimensions, multi-collinearity, and strong noise (Cao et al. (2022)). For example, the well-known decision tree algorithm recursively splits the training data based on the decision nodes. The optimal split is determined by maximizing a certain score function. The score function is sensitive to the training data. Some minor modifications to the original dataset result in an entirely different decision tree, which makes it difficult to generalize.

The decision tree is fairly easy to understand and implement. However, just one tree is not enough to produce valid results. A random forest consists of many decision trees, and there is no relationship between different decision trees. It randomly selects features for each decision tree, then averages the result (regression) or performs a majority voting (classification). A large number of uncorrelated decision trees will produce more accurate predictions than a single decision tree. But, the random forest model is prone to over-fitting due to the characteristics of industrial processes. Some studies have shown that random forests often overfit in the presence of noise (Biau (2012)).

To further avoid over-fitting, extremely randomized trees are proposed. ET is a tree-based ensemble method that uses a different type of decision tree compared to the random forest. It is superior to the random forest in terms of generalization and has outstanding performance when having redundant and noisy features. ET is similar to random forests but more robust and faster as it uses stronger randomization when splitting its decision tree node. We will describe the ET in detail in section 2.

Although tree-based ensemble models have achieved good results in many fields, little attention has been paid to explaining their predictions. These models have a common problem: the internal structure is very complex, which is difficult for humans to understand. The output of the model is also difficult to explain, which makes its application in some areas related to life safety or important decision-making very risky. Due to the risk-sensitive nature of industrial processes, the reliability and stability of soft sensors are essential for industrial applications. The ability to interpret soft sensor predictions can increase the reliability of soft sensors predictions. Therefore, it is crucial to understand the behaviour of the model and the important factors that affect the decision-making of the model through model interpretation (Du et al. (2019)).

Interpretable machine learning is a popular field of machine learning research (Murdoch et al. (2019)). An interpretable model is one that can estimate the contribution of each input feature to the model predictions. Shapley value is a concept based on the game theory proposed by economist Lloyd Shapley (Kuhn and Tucker (2016)). Its core idea is to fairly distribute the contributions of each player in a game, and then explain the black-box machine learning model from both global and local levels. If the Shapley value attribution is represented as a linear additive feature model, then it will be Shapley additive explanations (SHAP) model (Lundberg and Lee (2017)). It has a wide variety of applications as well as solid theoretical guarantees (consistency, local accuracy, and missingness) (Molnar (2020)). In this work, we use SHAP to explain ET-based soft sensor predictions, where the player is the input to the soft sensor, the game is the prediction of the soft sensor, and the SHAP value is the contribution of each input to the prediction.

This work aims to establish robust and interpretable industrial soft sensors based on extremely randomized trees and SHAP. The remaining part of this article is organized as follows. In Section 2, detailed explanations of extremely randomized trees and SHAP are given. Then, robust and interpretable inferential sensors are put forward with detailed implementation procedures. Section 3 presents a case study on the real fluid catalytic cracker (FCC) unit data to verify the effectiveness of the proposed method. Concluding remarks are presented in Section 4.

2. METHODS

2.1 Extremely Randomized Trees

Define S as $n \times p$ matrix, n as the number of training samples, p as the number of features, M as the number of trees, K as the number of features that are selected at each node, y as the output label and n_{min} as the minimum sample size for splitting a node.

ET is a tree-based ensemble method for supervised machine learning problems. Fig. 1 shows the structure of ET. To build ET, the first step is to create M decision trees.

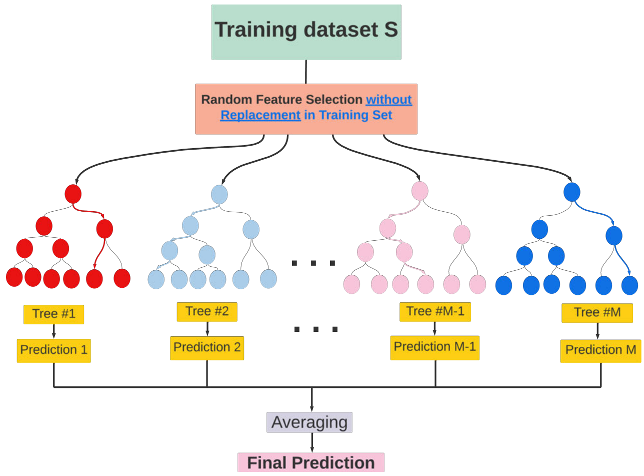

Fig. 1. Extremely Randomized Trees structure

Different from random forests, the sampling for each tree is random without bootstrap replacement. The usage of the full original training dataset (no bootstrap) can minimize the bias of ET. Then, K features among p features are selected randomly to develop ET. The value of K affects the randomness of the tree. In general, the smaller K is, the more random the tree is.

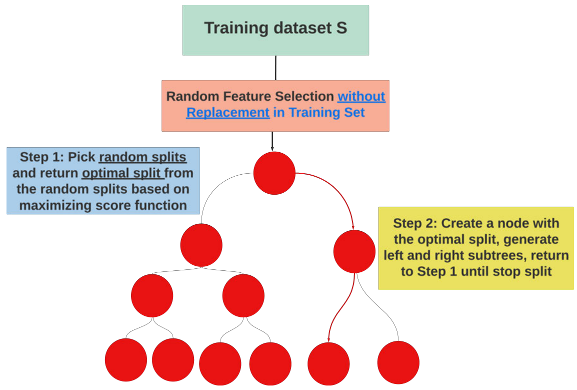

Fig. 2. Single tree structure of Extremely Randomized Trees

As we know, the traditional decision tree directly calculates the optimal split using entropy or information gain. Entropy is a measure of the uncertainty of a random variable and the information gain represents the degree of uncertainty reduction in the output y with a known feature X. The definition of entropy and information gain can be given as follows:

$$\begin{aligned} &P(X = x_i) = p_i, i = 1, 2, \cdots, n \\ &H(X) = -\sum_{i=1}^{n} p_i log p_i \\ &H(y \mid X) = \sum_{i=1}^{n} p_i H(y \mid X = x_i) \\ &IG(y, X) = H(y) - H(y \mid X) \end{aligned} \tag{1}$$

where p is the probability of random variable X, $H(X)$ is the entropy of random variable X, $H(y \mid X)$ is the conditional entropy of y given X, $IG(y, X)$ is the information gain.

Different from the traditional decision tree, every single tree of ET randomly selects the optimal split and divides the training data into subsets. They are done recursively until all training data subsets are correctly assigned or the sample size in subsets is smaller than n_{min}. Fig. 2 shows the structure of each tree in ET. The algorithm generates several random splits and then returns the optimal split (from these random splits) based on maximizing score function. Here, the score function in ET regression is defined as follows:

$$Q = \frac{var\{y \mid S\} - \frac{|S_l|}{S} var\{y \mid S_l\} - \frac{|S_r|}{S} var\{y \mid S_r\}}{var\{y \mid S\}} \quad (2)$$

where S_l and S_r represent the two subsets (left subset and right subset) of S that correspond to the split s, $var\{y \mid S\}$ is the variance of y in S. This score function is also called relative variance reduction. In fact, the construction of an entire tree is a method of dividing the space with a hyperplane. When compared to the optimal split (optimal hyperplane) utilized in RF that may cause overfitting of the original dataset, the use of random splits in ET provides greater diversity and robustness.

Fig. 3 gives the flowchart of an ET-based soft sensor. Table 1 summarizes the differences and similarities of decision trees, random forests, and extremely randomized trees.

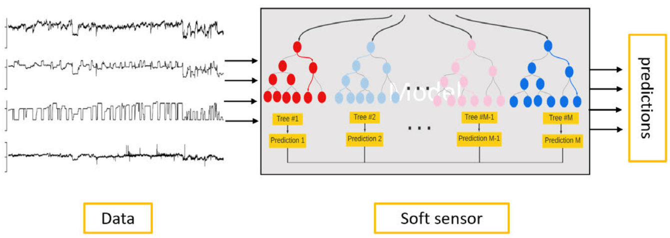

Fig. 3. An ET-based soft sensor

Table 1. Comparison of different tree methods

	ET	RF	DT
Number of trees	Many	Many	1
Decision node	Random features	Random features	All features
Split	Random split	Optimal split	Optimal split
Bootstrapping	No	Yes	NA
Variance	Low	Medium	High

The decision path of a tree is a straightforward interpretable approach, but for large-scale ensemble tree models, the decision path of each tree needs to be combined, and the results will become difficult to understand. Therefore, we need additional methods to interpret tree-based ensemble models.

2.2 SHAP (SHapley Additive exPlanations)

The Shapley value uses game theory ideas to assign feature contributions. Its main advantage is providing a consistent and fair solution. For the results predicted by multiple features, since there may be interactions between each feature, the Shapley value of a feature X is the weighted average contributions under all feature combinations.

The Shapley value of feature X is defined as follows:

$$\phi_X(f) = \sum_{S' \subseteq p \backslash X} w_X(S') \left[f\left(S' \cup \{X\}\right) - f(S') \right] \quad (3)$$

where f is the complex model like ET, $\phi_X(\bullet)$ is the Shapley value of feature X under model f, p is the number of input features, S' is a subset of the features. For $w_X(S') = \frac{|S'|!(p-|S'|-1)!}{p!}$, the denominator $p!$ represents all possible feature combinations; the numerator $|S'|!(p-|S'|-1)!$ means the appearance times of $S' \cup \{X\}$ in all $p!$ combinations; $f\left(S' \cup \{X\}\right) - f(S')$ indicates the expected marginal contribution of feature X in one combination.

When the Shapley value attribution is represented as a linear additive feature model, it is known as the Shapley Additive Explanations (SHAP) model. The SHAP model specifically adapts the Shapley values for interpreting the output of machine learning models. It quantifies the contribution of each feature to a particular instance's prediction while maintaining the desirable properties of Shapley values, such as consistency, local accuracy, and additivity. In the context of the SHAP model, the prediction can be decomposed into individual feature contributions and a baseline prediction (i.e., the prediction when no features are input), making it easier to understand the importance of each feature in a specific prediction.

From the definition of Shapley value, we can see that computing Shapley value is an NP-hard problem. In this work, a fast (polynomial-time) algorithm, TreeSHAP, is utilized to compute SHAP values for the ET-based soft sensor (Lundberg et al. (2018)). This is possible due to the structure of the tree-based model and the additivity of the Shapley values.

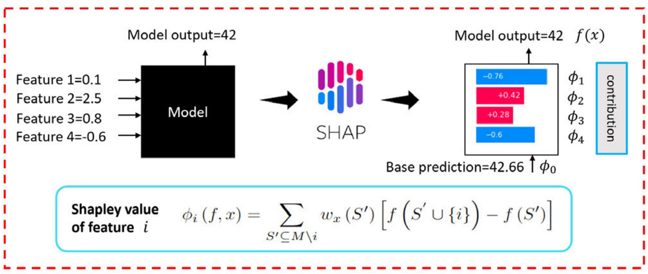

Fig. 4. An example of SHAP analysis

Fig. 4 shows an example of SHAP analysis, the black box model has 4 inputs and the output is 42. The base prediction is $\phi_0 = 42.66$. According to (3), the contribution of feature 1 is -0.76 and the contribution of feature 2 is +0.42, and so on and so forth. The sum of all individual contributions is equal to model output 42, which satisfies the definition of SHAP.

2.3 Interpretable Soft Sensors using Extremely Randomized Trees and SHAP

Tree-based interpretable models have significant implications for industrial process monitoring, as interpretation helps operators and engineers understand, trust and use

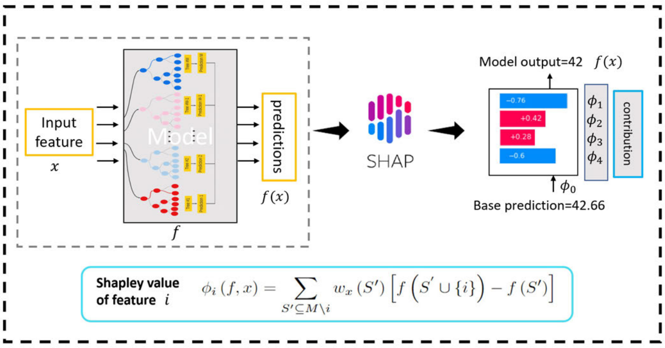

Fig. 5. The framework of proposed method

the model more effectively. In this work, we propose an accurate and interpretable soft sensor using ET and SHAP.

The framework of the proposed method is given in Fig. 5. The first step is data cleaning involving the removal of outliers and standardization of the cleaned data. Then, a robust and accurate ET soft sensor is developed with pretreated data. Finally, SHAP is used to accurately estimate the contribution of each input feature to the soft sensor predictions and the SHAP value is the contribution of the feature.

3. CASE STUDY

In this section, process data from the Parkland refinery in Burnaby, British Columbia, Canada, is used for the case study. We focus on establishing an interpretable soft sensor for a fluid catalytic cracker (FCC) unit. FCC is a core process in a refinery. It is an intermediate unit that processes the heavy hydrocarbons from crude oil and "cracks" them into smaller hydrocarbons, which can then be processed into a wide variety of different products (Su et al. (2021)). FCC unit consists of three main parts, namely the reactor, the regenerator, and the fractionator, which can be seen in Fig. 6.

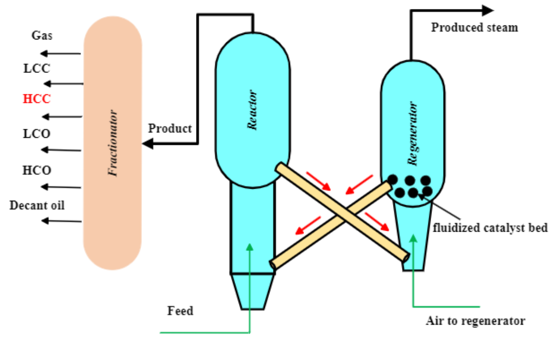

Fig. 6. A flow diagram of a Fluid Catalytic Cracking unit

The HCC (Heavy Catalytic Cracked) gasoline 90% cut point (the temperature at 90% volume distilled, cannot be measured online) in the FCC fractionator is selected as the soft sensor output. 10 process variables that may impact the cut-point temperature are selected based on process knowledge. We select 2076 samples from April 2018 to September 2022, of which the first 70% of the data (1453 samples) are used as the training set and the last 30% (623 samples) of the data are used as the test set. Considering confidentiality issues, we will not give the variable name and the magnitude of the variable. Fig. 7. shows the raw data after preprocessing.

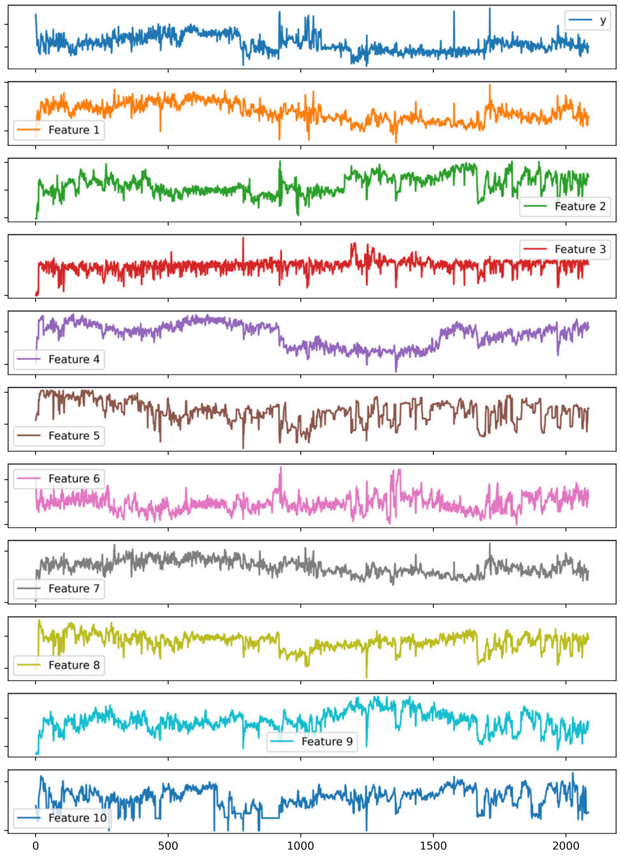

Fig. 7. Input and output data of the FCC unit

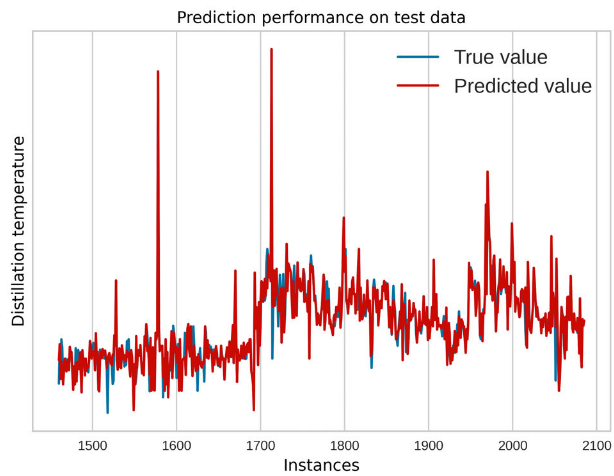

Fig. 8. The performance of ET soft sensor on test data

In this work, we use ET to construct soft sensors where the specific parameters, and the specific parameters were chosen with a grid search as follows: the number of trees M is 100, the selected number of features K at each node is 5, and the minimum sample size for splitting a node n_{min} is 2. Fig. 8 and Fig. 9 give the detailed prediction performance of the ET soft sensor on test data.

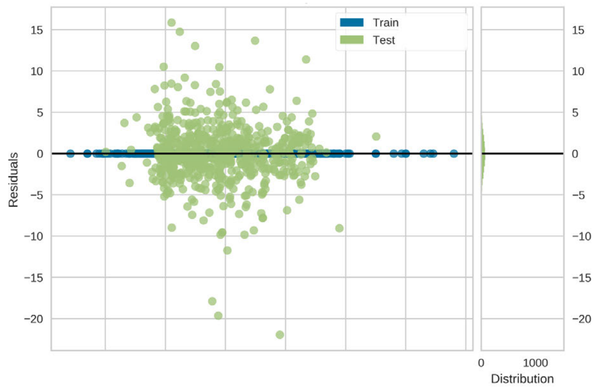

Fig. 9. The residuals of ET soft sensor on test data

Table 2. Comparison of different soft sensors

	RMSE	R^2
ET Regressor	3.8562	0.7932
Random Forest	4.0111	0.7771
Gradient Boosting Regressor	4.0301	0.7746
Huber Regressor	4.358	0.7367
Ridge Regression	4.4054	0.7311
Linear Regression	4.4067	0.7308
Neural networks (3 dense layers)	4.8609	0.6845
Lasso Regression	5.1631	0.6329
Elastic Net	5.3317	0.6093
Decision Tree Regressor	5.4937	0.5756

To evaluate the effectiveness of the proposed method, we compared it with several other machine learning methods, such as Random Forest Regressor, Gradient Boosting Regressor, Huber Regressor, Ridge Regression, Linear Regression, Neural Networks (3 dense layers), Lasso Regression, Elastic Net, and Decision Tree Regressor. We performed a grid search to select the optimal hyperparameters for each method and tested their performance on the test data. Table 2 summarizes the results of these experiments. The results prove that the proposed ET soft sensor has the best performance (RMSE is 3.8562, R^2 is 0.7932). It should be noted that ET soft sensor shows a larger improvement in performance compared to other tree-based methods, like Random Forest Regressor, Gradient Boosting Regressor, and Decision Tree Regressor, indicating that the proposed soft sensor is more accurate and robust.

Now we have a soft sensor of cut point temperature with excellent performance, but the problem is that the ET model itself has a complex structure, and it is difficult to know the inference process of the result from inside the model. Therefore, we use SHAP to enhance the interpretability of the model after training and mining the implicit information learned by the model.

SHAP can provide both global interpretation and local interpretation. Global interpretation refers to the interpretation of the entire model from input to output, from which we can understand the impact of each feature on the model. Fig. 10 displays the global interpretation of each input on the soft sensor prediction. Each instance is represented by a single dot on the feature row with the SHAP value (contribution) on the x axis. The sum of SHAP values is used to calculate the importance of the features, as shown on the y-axis. We can see that for all the data, the first and most important feature is feature 1; and for feature 1, the larger its feature value, the greater its contribution (positive correlation). Conversely, for feature 5, the smaller its feature value, the larger its contribution (negative correlation).

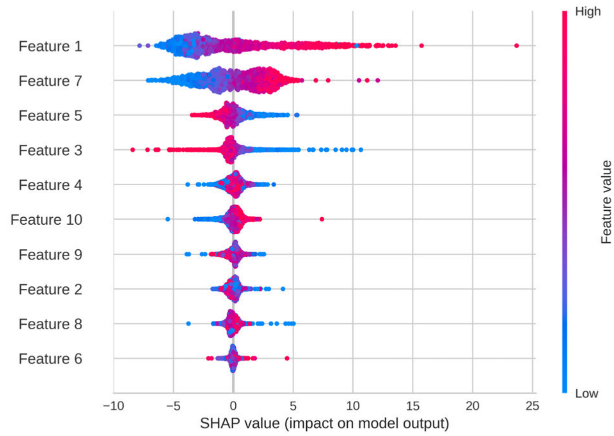

Fig. 10. Global interpretation of ET soft sensor

In addition to the global interpretation, we need to understand the variation in ET soft sensor predictions among specific instances. This type of explanation is called local interpretation. Local interpretation refers to explaining how the predictions change when the input values of an instance or a group of instances change. Fig. 11 shows the local interpretation of each instance on the ET soft sensor prediction. In this figure, $f(x)$ is the soft sensor prediction. For each individual prediction (column), the blue one means a negative contribution while the red one means a positive contribution. The darker the color, the greater the contribution.

Fig. 11. Local interpretation of ET soft sensor

To further show the interpretation of individual predictions, we choose the 1000th sample and the 2000th sample as examples for the analysis. The bottom of a waterfall plot starts as the base prediction (428.67), and then each row shows how the contribution of each feature moves the value from the base prediction to the ET soft sensor prediction.

As Fig. 12 shows, for the 1000th sample, the soft sensor prediction is 432.5. Feature 6 has the smallest contribution, moving only about 0.1 of the base prediction (428.67). Feature 5 has the largest contribution, moving about 3 of

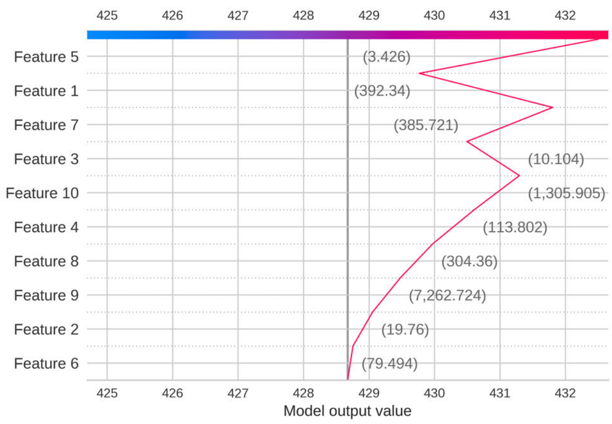

Fig. 12. Interpretation of ET soft sensor on 1000th sample

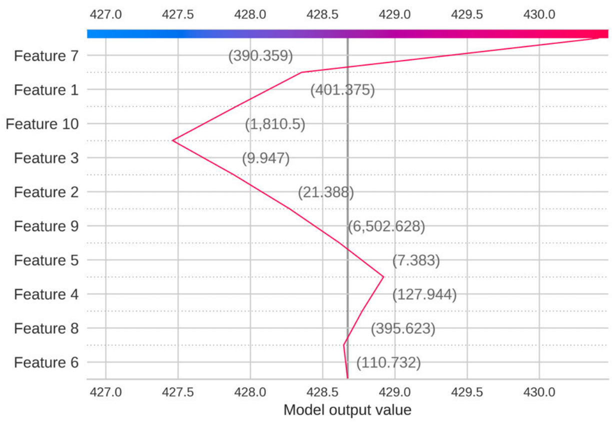

Fig. 13. Interpretation of ET soft sensor on 2000th sample

the base prediction (428.67). Similarly, as Fig. 13 shows, for the 2000th sample, the soft sensor prediction is 430.41. Feature 6 has the smallest contribution, moving only about -0.02 of the base prediction (428.67). Feature 7 has the largest contribution, moving about 2 of the base prediction (428.67). We notice a discrepancy between local and global interpretations. The reason for this discrepancy is that local interpretations focus more on specific instances, while global interpretations are concerned with the entire model. Furthermore, in specific instances, interactions between features may result in some features' contributions differing from those in the global interpretation. Therefore, we need to combine global and local interpretations to gain a more comprehensive understanding of the model's predictions.

4. CONCLUSION

The objective of this work is to make process monitoring methods more robust, efficient, and interpretable. Due to the unique characteristics of industrial processes, we introduce the ET algorithm to build an accurate and robust soft sensor. By increasing the randomness in the modeling process, ET solves the overfitting to a certain extent. In addition, since the ET model is not interpretable, SHAP is used to interpret complex ET model from global and local perspectives. The proposed explainable soft sensors using ET and SHAP can greatly improve interpretability while maintaining high accuracy. The effectiveness of the proposed interpretable soft sensor is demonstrated with a real application to a commercial-scale FCC unit.

ACKNOWLEDGEMENTS

We would like to thank Mitacs and Parkland Corporation for their financial support. We gratefully thank colleagues at Parkland for providing data and computing resources.

REFERENCES

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1), 1063–1095.

Cao, L., Su, J., Wang, Y., Cao, Y., Siang, L.C., Li, J., Saddler, J.N., and Gopaluni, B. (2022). Causal discovery based on observational data and process knowledge in industrial processes. *Industrial & Engineering Chemistry Research*, 61(38), 14272–14283.

Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3–42.

Gopaluni, R.B., Tulsyan, A., Chachuat, B., Huang, B., Lee, J.M., Amjad, F., Damarla, S.K., Kim, J.W., and Lawrence, N.P. (2020). Modern machine learning tools for monitoring and control of industrial processes: A survey. *IFAC-PapersOnLine*, 53(2), 218–229. 21st IFAC World Congress.

Ho, T.K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kotsiantis, S.B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261–283.

Kuhn, H.W. and Tucker, A.W. (eds.) (2016). *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, Princeton.

Lundberg, S.M., Erion, G.G., and Lee, S.I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.

Su, J., Cao, L., Lee, G., Tyler, J., Ringsred, A., Rensing, M., van Dyk, S., O'Connor, D., Pinchuk, R., and Saddler, J.J. (2021). Challenges in determining the renewable content of the final fuels after co-processing biogenic feedstocks in the fluid catalytic cracker (fcc) of a commercial oil refinery. *Fuel*, 294, 120526.

Zhu, Q., Joe Qin, S., and Dong, Y. (2020). Dynamic latent variable regression for inferential sensor modeling and monitoring. *Computers & Chemical Engineering*, 137, 106809.