Computers and Chemical Engineering 162 (2022) 107795

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

A novel zone-based machine learning approach for the prediction of the performance of industrial flares

Helen H. Lou[a,*], Jian Fang[a], Huilong Gai[a], Richard Xu[b], Sidney Lin[a]

[a] Dan F. Smith Department of Chemical and Biomolecular Engineering, Lamar University, 4400 MLK Blvd., PO Box 10053, Beaumont, TX 77710, USA
[b] Clear Lake High School, 2929 Bay Area Blvd., Houston, TX 77059, USA

ARTICLE INFO

Article history:
Received 24 October 2021
Revised 25 March 2022
Accepted 30 March 2022
Available online 8 April 2022

Keywords:
Flare performance prediction
Random forest
Catboost
Zone-based model

ABSTRACT

Industrial flares are used to burn off unwanted gas during operation. If not combusted completely, intermediate products or incomplete combustion products are formed, and they will cause significant environmental and health issues. The EPA Refinery Sector Rule emphasizes smokeless flaring with combustion efficiency (CE) ≥ 96.5% and destruction and removal efficiency (DRE) ≥ 98% for all types of flares in the refineries. In this research, a novel zone-based modeling approach was developed for predicting CE and Opacity of steam assist flares. Flare CE data were partitioned into two zones based on the partition of the carbon and hydrogen atomic ratio (CHR), then random forest (RF) and Catboost algorithms were used to develop CE predictive models, respectively. This CHR-based zone partition has a clear implication in engineering. It was also found out that no zone division for flare Opacity prediction is needed, and both RF and Catboost algorithms generated good prediction results. All the models match extremely well with all the original experimental data. These predictive models under the same zone-partition use either RF or Catboost algorithm can both give superior prediction accuracy. This demonstrates the simplicity, general applicability, and high reliability of the zone-based ML approach.

1. Introduction

Currently, the industry is going through "The Fourth Industrial Revolution", also called Industry 4.0, which is featured by the integration between physical and digital systems of production environments (Carvalho et al., 2019). The availability of massive amount of data has prompted many industries to reposition themselves to take advantage of the disruptive potentials of data analytics and machine learning (ML) (Abubakirov et al., 2020). The rapid and successful development of ML, combined with the advent of the Internet of Things, Big Data, and the fourth industrial revolution, have transformed many industries, including energy industries and chemical process industries (CPI).

Flares are important safety devices for pressure relief from process units, oil and gas fields, pipelines, and storage vessels during normal operation as well as start-up, shutdown, and malfunction (SSM) situations. Flaring can dispose flammable "waste" gases in downstream refineries and chemical facilities, midstream, and upstream industries. However, during flaring, undesirable pollutants, including unburned hydrocarbons (such as methane) and combustion byproducts like soot, CO, CO_2, unburned volatile organic, and oxides of nitrogen and sulfur, are formed from the combustion process (Singh et al., 2014). These emissions cause significant health and environmental impacts.

In 2015, the Refinery Sector Rule (RSR), 40 CFR 63.670, was published by the U.S. Environment Protection Agency (EPA), which requires smokeless flaring and aims at achieving the objectives of 96.5% combustion efficiency (CE) and 98% destruction and removal efficiency (DRE). Currently, RSR assumes that CE and DRE objectives and smokeless flaring are achieved if the flares are operated under specific conditions: Net Heating Value Combustion Zone (NHV_{CZ}) > = 270 BTU/scf, or Net Heating value dilution parameter (NHV_{dil}) > = 22 BTU/ft^2 and Vent Gas Velocity (V_{tip}) < V_{max} & V_{tip} < 400 ft/s (US EPA, 2016). CE denotes the percentage of hydrocarbon in the flare vent gas that is completely converted to CO_2 and water vapor. DRE denotes the percentage of a specific pollutant in the flare vent gas converted to a different compound (such as CO_2, CO, or other hydrocarbon intermediates). DRE of a flare will always be greater than CE. It is generally estimated that a CE of 96.5% is equivalent to a DRE of 98% (US EPA, 2016). However, flaring performance depends on many parameters, such as the compo-

* Corresponding author.
E-mail address: Helen.lou@lamar.edu (H.H. Lou).

https://doi.org/10.1016/j.compchemeng.2022.107795
0098-1354/

sition, flow rate, velocity, and heating value of the vent gas, wind speed, the flow rate of assist steam or air, and makeup fuel. Besides, Flare operators should establish the smokeless capacity to ensure 98% DRE or 96.5% CE or higher at all times and assess the exceedance of the smokeless capacity based on cumulative volumetric flow rate and/or flare tip velocity (US EPA, 2006). Still, there is no guarantee that the CE will be $\geq$ 96.5% and DRE will be $\geq$ 98% all the time (Zeng et al., 2016). All these issues lead to questions of how to operate the flares in a most environmentally responsible and cost-effective manner to comply with the regulations, achieve smokeless flaring, or even maximize CE and DRE under a given set of design and operating conditions. A few studies were performed to predict flare emissions and performance using data-driven models as presented below.

While the contributions of chemical engineers in process monitoring and inferential sensors are well known, thousands of papers have been published in recent decades using latent variable methods, neural networks, and other ML or statistical methods in CPI (Qin and Chiang, 2019). Lou and Gai (2020) provided of a novel method of "trustworthy AI (TAI)" for the CPI and its successful application in two chemical processes for improving operational excellence. Recent advances include the development of support vector machines (SVM) and kernel methods (Nguyen et al., 2021). Statistical ML becomes a major framework for artificial intelligence, which bridges computations and statistics with information theory, signal processing, control theory, and optimization theory (Qin and Chiang, 2019).

Complex models using deep neural networks can give accurate prediction in many circumstances. On the other hand, in order to achieve interpretability, simpler models are preferred (Qin and Chiang, 2019). Interpretability is critical for industrial adoption since operators and decision-makers must establish trust in the algorithms. For practical application in the industry, it is encouraged to build models that are more accurate but simpler.

Ensemble learning is a branch of machine learning, and it is mainly divided into Bagging and Boosting algorithms (Gomez et al., 2010; Prokhorenkova et al., 2019). Bagging renders the model a high generalization by reducing the variance. Random Forest (RF) is a popular Bagging-based ensemble learning algorithm for regression or classification (Breiman, 2001; Chrysafis et al., 2017; Louppe, 2014; Scornet, 2015). It is a meta estimator that fits a number of decision trees on various sub-samples of the dataset and uses majority voting (for classification problems) or averaging (for regression problems) to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn randomly. The predictions from the forests are averaged using bootstrap aggregation and random feature selection. RF models have been demonstrated to be robust predictors for both small sample sizes and high dimensional data (Biau and Scornet, 2016). Smarra et al. (2018) built a Data-driven model of Predictive Control (DPC) based on historical building data using a regression tree and RF algorithm and applied DPC to three case studies to demonstrate the performance, scalability, and robustness. Zimmerman et al. (2018) explored the appliance of RF calibration models to air quality. Wang et al. (2019) presented a novel data-centric predictive control (DPC) approach which utilizes RF for online control and optimization of a nonlinear chemical process, and reported a case study of feed rate control for aluminum smelter cells.

Catboost, an adaptive boosting ensemble learning algorithm based on decision trees, is powerful and very fast (Hancock and Khoshgoftaar, 2020). It is widely applied to multiple data types to solve a wide range of problems like fraud detection, recommendation items, and forecasting. Catboost can also return very good results with relatively less data, unlike the classical algorithms that need to learn from a massive amount of data (Luo et al., 2021). Hancock and Khoshgoftaar (2020) developed an hourly forest fire risk index (HFRI) with 1 km spatial resolution using accessibility, fuel, time, and weather factors based on Catboost machine learning. The performance of HFRI ensemble model was better than the meteorological model. The Catboost model confirmed that most forest fires were caused by anthropogenic factors. He et al. (2021) combined a Catboost model with a mechanism model and applied it to an industrial fluid catalytic cracking unit to maximize the yield of iso-paraffins. In the test, the total liquid yield was increased by 1.19% on average, and the coke yield was decreased by 1.05% on average. In particular, the computing time was reduced from more than 20 h to less than 1 min.

Machine learning is widely used to analyze large-scale data sets in this age of big data. Extracting useful predictive modeling from these types of data sets is a challenging problem due to their high complexity. One ML approach to handle a large data set is to partition the data set into subsets, run the learning algorithm on each of the subsets, and combine the results (Chan and Stolfo, 1995). Some ML approaches use weak learners in the partitioned data, and then build a strong learner from a set of weak learners (Çatak, 2017). For example, Li et al. (2022) developed prediction models for earthquakes using four ML algorithms based on 113 datasets, which were developed into different spatial divisions based on the partition of two features: the densities of population and geological faults. It was found that the predictive performance of different types of algorithms are significantly different under the same partition. Mohammadi et al. (2022) developed the advanced random forest (RF) and Boosted regression tree (BRT) models to understand the factors influencing surface water vulnerability to arsenic pollution. The data were partitioned into two zones by arsenic concentration in the sediment samples (polluted: $>$ 8 ppm and non-polluted: $<$ 8 ppm), 144 polluted arsenic sediment samples and 144 non-polluted arsenic sediment samples. These reported researches used prior domain expertise, for example, population and geological faults for the severity of earthquake damage, or known obvious facts, for example, arsenic concentration in the sediment determines surface water arsenic pollution. They did not address the issue of how to find the best feature to partition the data when there is no clear clue.

It is hypothesized by the authors that "one ML model first all" approach does not work well for complicated systems/processes. This study seeks to develop the applicable novel zone-based ML approach to identify the best zone partitions of the data and the corresponding predictive zone-based ML sub-models. This zone-based ML approach is not dealing with a well-known, well-explained situation that gives the modeler a clear direction on which feature should be used for zone partition. A set of novel zone-based ML models were successfully developed to predict flaring performance of steam-assisted flares. Random forest, representing Bagging algorithm, and Catboost, representing boosting algorithm, were utilized for comparison and to justify the robustness of the zone-based approach.

The unique contribution of this novel zone-based ML approach is that this algorithm is applicable when the modeler does not know which feature is the best one for data partition. This algorithm can figure out the best data partition approach automatically and develop high-performance ML models. This algorithm was applied to the prediction of the flaring performance. The novel zone-based ML approach eventually found out the carbon and hydrogen atomic ratio (CHR) is the best choice for data partition to predict combustion efficiency. While for Opacity prediction, a one-piece ML model is sufficient. The zone-based models for CE and Opacity predictions have significantly higher predictive accuracies than those of the prior one-piece models, and none of the original data were deleted.

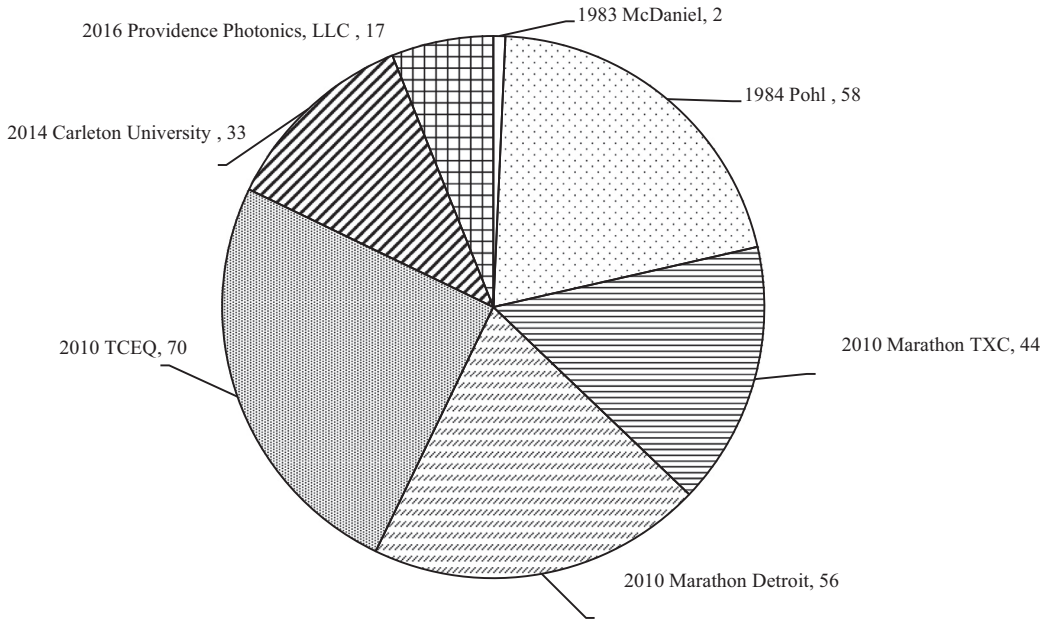

Fig. 1. Data source of the flare experimental data.

2. Data collection

The flare experimental data mainly come from the following resources: flare efficiency study completed by EPA in 1983 (McDaniel and Tichenor, 1983), evaluation of the efficiency of industrial flares (Pohl et al., 1984), Texas Commission of Environmental Quality (TCEQ) 2010 flare study final report (Allen and Torres, 2011a, 2011b), Marathon Petroleum company flare study reports (Cade and Evans, 2010; Ewing et al., 2010), Carleton University soot emission rate measurement results (Corbin and Johnson, 2014; Johnson, 2014), and Providence Photonics, LLC (Zeng et al., 2016). The distribution of the data is shown in Fig. 1.

The data collected from the literature include the geometry of the steam-assist flares represented flare tip diameter ("D"), operating data such as net heating value of combustion zone ("NHV_{cz}"), net heating value of vent gas ("NHV_{vg}"), flare tip exist velocity ("V"), steam assist flow rate ("S"), ratio of actual assist steam to stoichiometric steam ("SER"), ratio of carbon and hydrogen atomic ("CHR"), molecular weight of vent gas ("MW"), carbon number of fuel species ("CN"), flare efficiencies ("CE"), soot emission ("Opacity"), p-bond in vent gas ("Pi_{vg}"), and p-bond in the combustibles ("Pi_{cz}"), and the meteorological data including crosswind speed ("U") and crosswind speed/flare tip exit velocity ("U/V"). A total of 280 sets of the flare experimental data are available. The composition of the flare gas includes methane, ethane, ethylene, propane, propylene, etc. In those experiments, no make-up fuel was used since the net heating values of the vent gas were high.

Fig. 2 shows box plots of the flare experimental data consisting of 280 datasets for 13 dependent variables (features) and two independent variables (model output: CE and Opacity). In Fig. 2, it is clear that data for all 13 features follow a normal distribution, and the outliers of each feature are less than 0.7% of the total data available. In addition, the model input (NHV_{vg} and NHV_{cz}) data are more dispersed, covering a wide range. Thus, the quality of the database is acceptable for developing good ML models.

In Fig. 3 (a), the histogram of CHR showed that the CHR spread is from 0.215 to 0.504 and its distribution skewed left. The data mainly distributes in the bins of 0.30 – 0.35, 0.35 – 0.40, and 0.45 – 0.504. The scatter plot in Fig. 4 did not disclose any obvious zone-based relationship between CE and CHR. However, the zone-based ML approach eventually figured out CHR is the best feature for zone partition in CE prediction.

Since the composition of the flare gas includes different fuel mixtures, such as methane, ethane, ethylene, propane, propylene, etc., features including Pi_{vg}, Pi_{cz}, CN, and CHR of vent gas were introduced to account for the effect of vent gas composition (Alphones et al., 2020).

CHR reflected different species of gas in the combustion zone. Alkanes or alkenes include the sigma bonds, the double bonds or triple bonds. A single bond is a sigma bond that is formed by ending the overlap of two atomic orbitals, hybrid orbitals, or one of each (Albright et al., 2013). The first bond between any two atoms is always a sigma bond and any additional bonds are pi bonds that stand for the overlap of two adjacent parallel p orbitals. Thus, a double bond has one sigma and one pi bond. A triple bond has one sigma and two pi bonds. Additionally, alkenes and alkyne with higher CHR have p-bonds, which are highly correlated with SP^2 hybridization of carbon atoms, and may also lead to higher Opacity, as absorption efficiency increases due to high electron density (Jäger et al., 1999). Prior research found CHR is more suitable for flare performance prediction than Pi_{vg}, Pi_{cz}, and CN (Wang, 2019).

Note that the CHR values for CH_4, C_2H_6, C_2H_4, and C_3H_8 are 0.250, 0.333, 0.500, and 0.375, respectively. Alkenes and alkynes with higher CHR consume more oxygen, which is highly correlated with the net heating value of combustion zone and net heating value of vent gas, may lead to lower CE and higher Opacity as the consumption of oxygen increases due to insufficient air assist flow rate. This is consistent with the conclusions of Trivanovic (Trivanovic et al., 2020), which showed that flare gas with higher heating value produced substantially more soot.

3. Prior research

It is stated that the flare.IQ Advanced Flare Control Solution by Baker Hughes ensures high-efficiency flare combustion, while reducing methane emissions and steam usage in flare systems. The flare.IQ software is pre-programmed with the company's patented algorithms to calculate molecular weight. From molecular weight, the net heating value of the flare gas is determined. Fuel gas demand and steam demand are then set based on current net heat-

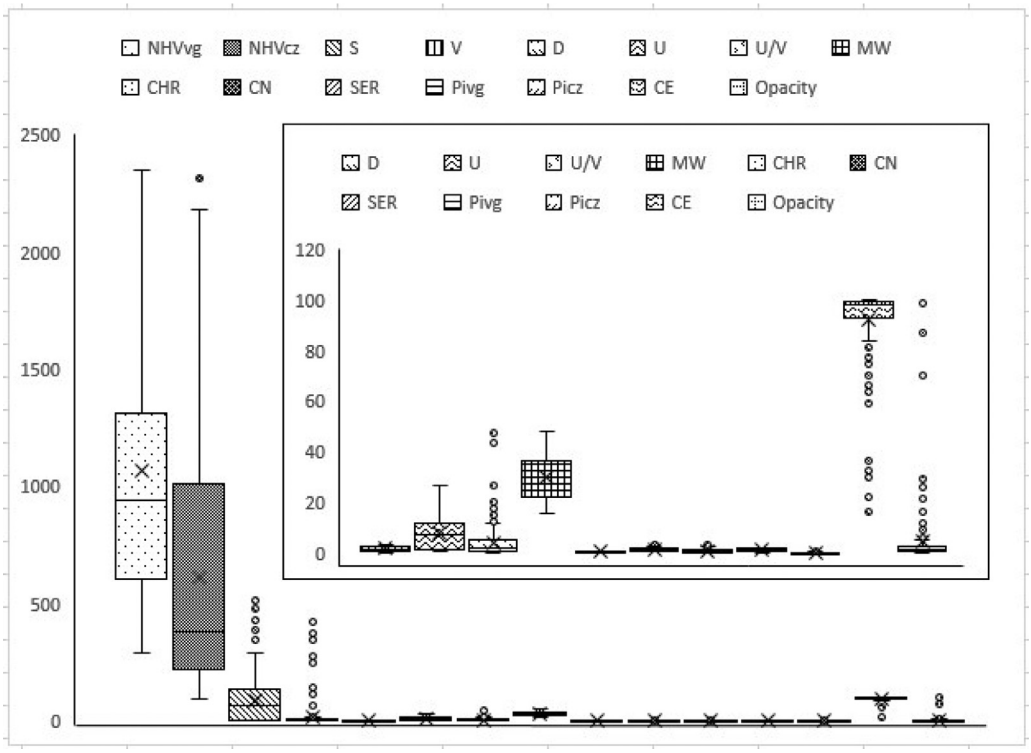

Fig. 2. Box-plots of the flare experimental data.

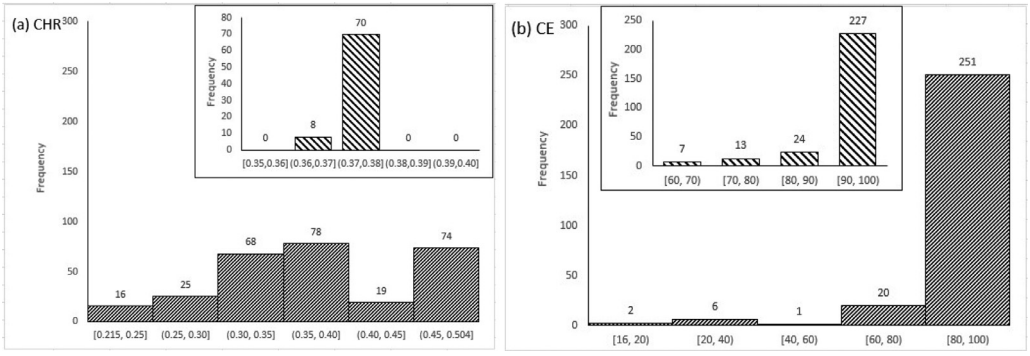

Fig. 3. Frequency histograms of the CHR (a) and CE (b).

ing values, hydrocarbon molecular weight, and flare vent gas flow (Hughes, 2019). It's stated that the flare.IQ is capable of operating with a wide variation of flare installations but its core algorithms are not disclosed.

In Lamar University's flare research team, the predictive models based on the experimental data have been systematically developed using three algorithms: response surface model (Alphones et al., 2020), artificial neural networks (Damodara et al., 2020), and random forest algorithm (Wang, 2019).

3.1. Response surface models

The robust response surface models were developed using Minitab 18 statistics toolbox to express CE and Opacity as a function of operating variables for the experimental data, and then the models were validated based on R^2 (Alphones et al., 2020). Outliers were removed in each step based on standardized residual plot analysis. The adequacy and significance of each parameter in the models was checked by the analysis of variance table and p-value, respectively. It was found that CHR, CN, NHV_{cz}, U/V, S, and U had significant impacts on Logit (100 - CE), while CHR, CN, NHV_{cz}, V, S, and U had significant effects on Logit (Opacity) for steam-assisted flares.

The quadratic form of the response surface model, General Quadratic Response Surface Models, or GQM is expressed as

$$y = \sum_i a_i x_i + \sum_i a_{ii} x_i^2 + \sum_i \sum_j b_{ij} x_i x_j + C + e \quad (1)$$

where y is the independent variable, x is the dependent variable, i and j are numbers from 1 to n (n are the number of variables), a_i are the linear coefficients, a_{ii} are the quadratic coefficients, and b_{ij} are the interaction coefficients, C is the constant term, e is the residual error.

Seventy seven (77) data were considered as outliers for CE. 80% of 203 data were randomly selected using Excel for GQM development, and the remaining 20% of data were used for GQM validation, such that the data used in model validation lie within the range of the variables used in models. The R^2 value of the predictive GQM-CE was 0.93, and its mean absolute error (MAE) value was 2.1 for steam-assisted flares, and the mathematical function

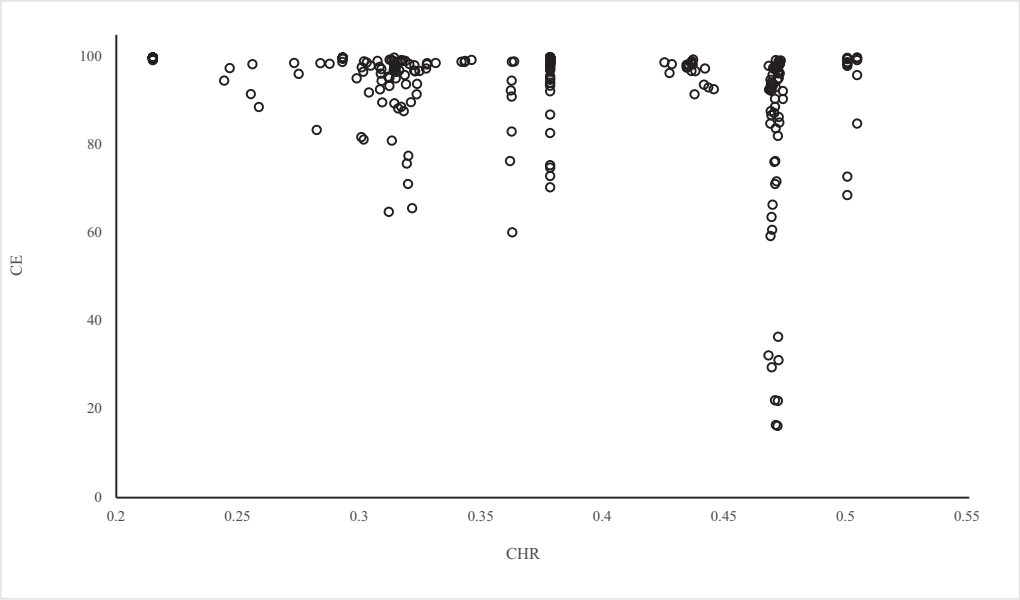

Fig. 4. Scatter-plot of CHR vs. CE.

was derived as shown below:

$$\begin{aligned} Logit\ (100-CE) &= 0.237 - 7.06\ \times 10^{-3} \times NHV_{CZ} + 1.88\ \times CN \\ &+ 1.6\ \times 10^{-6} \times NHV_{CZ}^2 -\ 8.7\ \times 10^{-6} \times S^2 - 12.3 \times CHR^2 \\ &+\ 3.3\ \times 10^{-3} \times U/V \times U - 3.68\ \times 10^{-4} \times U/V \times S \\ &- 2.9\ \times 10^{-5} \times U \times NHV_{CZ} - 6.6\ \times 10^{-5} \times NHV_{CZ} \times S \\ &+ 9.3\ \times 10^{-3} \times NHV_{CZ} \times CHR - 9.28\ \times 10^{-4} \times NHV_{CZ} \times CN \\ &+ 0.054\ \times S \times CHR - 6.77\ \times 10^{-3} \times S \times CN \end{aligned} \tag{2}$$

Ninety six (96) data were considered as outliers for opacity. 80% of 184 data were randomly selected using Excel for GQM development, and the remaining 20% of data were used for GQM validation, the R^2 value of the predictive GQM-Opacity was 0.91, and its MAE value was 0.94 for steam-assisted flares, and the mathematical function was derived as shown below:

$$\begin{aligned} Logit\ (Opacity) &= -2.36 - 2.4\ \times 10^{-3} \times NHV_{CZ} - 8.1\ \times 10^{-3} \times S \\ &- 0.034\ \times V + 0.0777\ \times U + 1.67\ \times CN + 1.4\ \times 10^{-5} \times S^2 \\ &+ 4.4\ \times 10^{-5} \times V^2 - 3.02 \times U^2 -\ 0.3774\ \times CN^2 \\ &+ 6.7\ \times 10^{-5} \times NHV_{CZ} \times V + 3.5\ \times 10^{-3} \times NHV_{CZ} \times CHR \\ &+ 6.6\ \times 10^{-4} \times CN \times NHV_{CZ} + 5.4\ \times 10^{-4} \times V \times S \\ &- 5.7\ \times 10^{-3} \times S \times CHR \\ &- 2.4\ \times 10^{-3} \times V \times U - 0.039\ \times V \times CN \end{aligned} \tag{3}$$

3.2. ANN algorithm

Based on the GQM, ANN models were developed for the flares experimental data. The ANN model consists of a two-layer feedforward network using one hidden layer of 'tansig' neurons followed by an output layer of linear neurons in the neural network toolbox in MATLAB (Damodara et al., 2020). The network was trained using the Levenberg-Marquardt backpropagation algorithm. It was found that seven important variables (CHR, CN, NHV_{cz}, U/V, S, U, and D) and five important variables (NHV_{cz}, U/V, S, U, and D) were used to develop two Logit (100 - CE) models, and three important variables (CHR, NHV_{cz}, and U/V) and two important variables (NHV_{cz}, and U/V) were used to build up two Logit (Opacity) models. The ANN predictive models have three layers (input, middle, and output). The middle layer for CE prediction has 7 neurons while the middle layer for Opacity prediction has 5 neurons. They were validated based on R^2 and MAE values.

The high accuracy of ANN models for CE and Opacity confirmed the utility of ML algorithm. However, no systematic feature selection was conducted, and the ANN predictive models were complex. These limit the potential implementation of ANN models for online application.

3.3. Random forest (RF) algorithm

In Wang's work (Wang, 2019), based on the aforementioned experimental data, the following seven features, NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V, were chosen as the top features for CE and opacity predictive models. Comparing with NHV_{vg}, NHV_{cz} considers steam, air, makeup fuel added into the combustion zone. High-speed assist steam can also draw ambient air into the combustion zone and help to increase the mixing of the steam and air with vent gas exiting the flare tip (Allen and Torres, 2011a). However, the assist steam had a negative effect on CE. Fifty five (55) outliers were removed using local outlier factor and isolation forest methods. For the remaining 225 data, 90% of data (202) were randomly chosen as training data and the remaining 10% of data were used to test the prediction RF models, respectively. RF predictive models of CE and opacity were built. The performances of the RF models were evaluated using R^2 and MSE values. The R^2 values of the predictive RF-CE and RF-Opacity models were 0.992 and 0.936, and their MSE values were 1.320 and 0.980, respectively.

4. A novel zone-based ML approach

In statistics, an outlier is a data point that differs significantly from other observations (Maddala, 1992). Even though the so-called "outliers" computed by the ML partition algorithms have a clear statistical implication, the exclusion of too many "outliers" may significantly reduce the amount of available data, which hinders the general applicability of the models. In an industrial set-

ting, it is true that some data points differ significantly from the other data points due to process variation, sensor malfunction, etc. However, sometimes the "outliers" issue arises because the model does not fit the data very well. In order to increase model accuracy, the modeler may declare some "bad data" as "outliers" and delete them to increase model accuracy. On the other hand, if a model is generally applicable, the issue of "outliers" may be avoided. It is hypothesized by the authors that "one ML model first all" approach may not work well for complicated processes and may cause the unnecessary issue of "outliers". In our prior research of flaring performance prediction based on reported experimental data, some ML models with high accuracy were obtained after deleting a large number of "outliers". A synopsis of the group's prior research is provided in Section 3. However, all the reported flaring experiments were well planned and well conducted by the industry, researchers, and/or environmental protection agencies, following strict quality control protocols. Therefore, the authors hypothesized that these "outliers" are not bad data, they are just not fit for a one-piece ML model. Since the combustion reactions during the flaring process are different under different conditions, we can characterize different flaring processes using different models. A novel zone-based ML modeling approach ensuring model accuracy and avoiding deleting outliers, was developed in this study.

4.1. The general methodological framework

A novel zone-based ML modeling approach was introduced in this study. The basic concept of the approach is the division of the whole dataset into several distinct zones, instead of identifying and removing outliers. This was especially important for predictive flaring performance in the real-world applications since flare CE was strongly dependent on flare gas composition. The idea was the replacement of a narrowly applicable, low accuracy model by a high accuracy, unified zone-based model. However, the development of zone-based ML models is challenging, since it involves the identification of boundary points for each zone, partitioning the input and output space, and optimization of each sub-model.

The general methodological framework for a novel zone-based ML approach is presented in Fig. 5. In the "I: Data Preparation" part, during the "Data Cleaning" step, the collected data are cleaned, including converting data types, filling incomplete data, fixing or removing incorrect data, and identifying "outliers". Then the modeler will decide if the "outliers" are true outliers, or a zone-based ML model is more suitable to accommodate these outliers by developing and comparing the performance of one-piece models vs. zone-based models. When developing the one-piece models, the "outliers" identified in the "Data Cleaning" step are deleted; When developing the zone-based models, the "outliers" are kept. In the "II: Zone-based ML Model" part, the algorithm utilizes a loop to identify the best feature for the data partition into different zones. Then, this algorithm generates the best zone-partitions automatically, builds the corresponding zone-based ML models, evaluates them and identifies the best zone-based ML models. If the best zone-based model performs better than the best one-piece model, then the formal one will be the final choice.

4.2. Best feature for zone partition

Feature importance and the best feature for zone partition are fundamentally different. Feature importance calculates a score for all the input features for a given model — the scores simply represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. The important features can be identified through the feature selection methods, such as Pearson correlation coefficient and Feature dependency matrix, etc. (Benesty et al., 2009; Ye and Liu, 2005). Our prior research showed that for a one-piece ML model, the most important features are NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V (Wang, 2019), and their feature importance results are shown in Table 1.

Table 1
Features importance results for the prediction of CE.

Features	Mean decrease impurity importance
NHVcz	0.3837
S	0.1954
NHVvg	0.1627
U/V	0.1074
U	0.0931
CHR	0.0442
D	0.0136

However, feature importance is calculated based on a one-piece model and feature importance can not tell which feature is the most important one for zone partition. This is a fundamental flaw in the typical workflow of data analytics and machine learning. If we use the feature importance for zone partition, the choice should be NHV_{cz}, and nobody will consider CHR. However, our algorithm eventually found out CHR is the best choice in CE prediction.

In refineries and combustion plants, the composition of the vent gas changes all the time, which means the chemical properties of the net heating value of vent gas (NHV_{vg}) and CHR change all the time. The fluctuation in NHV_{vg} affects NHV_{cz}. Trivanovic (Trivanovic et al., 2020) showed that flare gas with a higher heating value produced substantially more soot. Additionally, alkenes and alkyne with higher CHR have p-bonds, which are highly correlated with SP^2 hybridization of carbon atoms, and may also lead to higher Opacity, as absorption efficiency increases due to high electron density (Jäger et al., 1999). This research found out that both NHV and CHR are important for the flaring performance but did not give a conviction which feature or the combination of these features are the best for data partition. USEPA uses NHV_{cz} to guild the flare operation and CHR is not mentioned in the guideline (US EPA, 2016). So the common sense is that NHV_{cz} is the most influential feature for flare combustion.

This zone-based ML approach is not dealing with a well-known, well-explained situation that gives the modeler a clear direction on which feature should be used for zone partition. Instead, this algorithm figured out the best data partition approach automatically and give a unique contribution.

4.3. A zone-based ML approach

At the beginning of ML modeling work, it is not known how many zones need to be divided for the i^{th} feature ($1 \le i \le I$) and how to identify the boundary points for zone k ($1 \le k \le K$). The pseudocode of the zone-based ML approach is shown in Fig. 6, and the logic flow of the algorithm is depicted in Fig. 7.

Step 1: Set up the outer loop to investigate how many zones need to be divided for a feature. In the outer loop, each feature can be randomly selected, or the modeler can investigate the zoning potential of each feature based on its importance to the target output, or based on some scientific principles.

Step 2: Set up the middle loop to identify the number of zones. At the beginning, the datasets are sorted in an ascending order of the current feature. Then, the lower boundary point of zone (LBZ) for zone 1 is set at the minimum value of feature i ($X_{i,MIN}$). The upper boundary of the zone (UBZ) for zone k ($1 \le k \le K$) is the target to be identified in future steps.

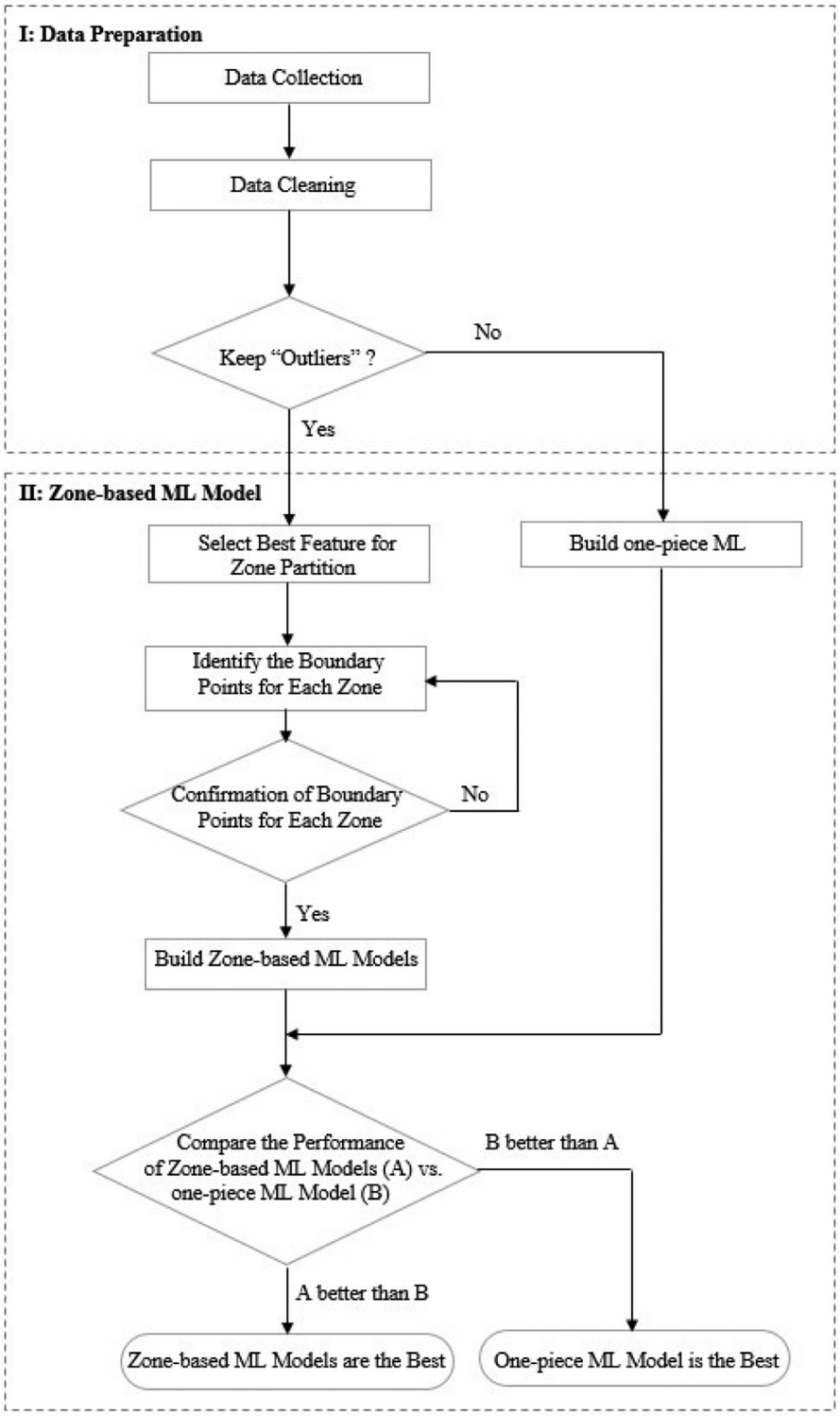

Fig. 5. The general methodological framework for a zone-based ML approach.

Step 3: Build the inner loop to identify the best ML predictive model in each sub-zone. This task can be accomplished by testing the performance of the ML model throughout the data in each sub-zone. The model evaluation criterion can be set up based upon the modeler's preference.

Step 4: Figure out the UBZ of zone k, which is also the LBZ of zone k - 1. For example, the UBZ of zone 2 is identified by repeating Step 3. The middle loop iteration repeats until the value of the UBZ of zone k is greater than the maximum value of feature i ($X_{i,MAX}$), at which point the command flow returns to the outer loop.

Step 5: Investigate how many zones need to be classified for the rest features by repeating Steps 2, 3, and 4. The outer loop iteration repeats until all the features have been processed, and the command flow ends. Finally, the key feature, the partition of zones, and the best zone-based ML predictive sub-models are iden-

```
START
for i = 1 to data-features.size step 1 do
    identify every feature X_i
    sort original data in ascending order of X_i
    let Y_{i,[1...a]} be a new submatrix
    for j = 1 to data-row.size step 1 do
        identify the minimum number of zones k
        Y_{i,a} ← X_{i,j} in each zone
        for l = 1 to k step 1 do
            identify the best ML model using the submatrix Y_{i,a}
        end for
        print zone-based ML model in each zone
    end for
    print zone classification
end for
print key feature
STOP
```

Fig. 6. The pseudocode of the zone-based ML approach.

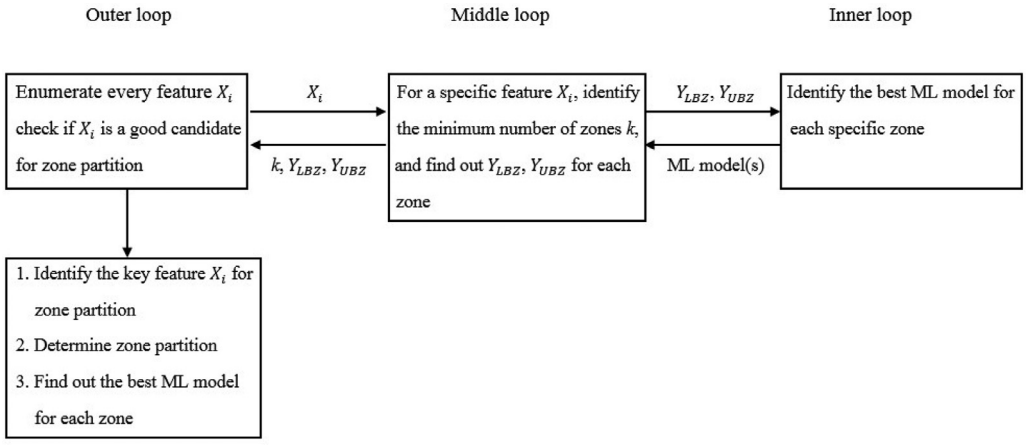

Fig. 7. Detailed breakdown of nested loops in the zone-based ML approach.
Note: Y_{LBZ}: the lower boundary point of the current zone (LBZ); Y_{UBZ}: the upper boundary point of the current zone (UBZ).

tified. In this case, the criterions are $R^2 \geq 0.99$ and the minimal MSE value.

4.4. Identification of boundary points for each zone

After zone division, the sub-models for each zone can be developed using any suitable ML algorithms. In this work, two ensemble learning algorithms, random forest (representing the bagging algorithms) and Catboost (representing the boosting algorithms) were chosen.

For CE prediction, this research was found out that CHR is the best feature for zone partition of the entire datasets. The values of CHR ranges between 0.215 and 0.504 ($X_{CHR,MIN} = 0.215$ and $X_{CHR,MAX} = 0.504$), while the values of available CE ranges from 16.00 to 100.00, expressed as $Y_{CE,MAX} = 100.00$ and $Y_{CE,Min} = 16.00$, as shown in Fig. 3 (a) and (b), respectively. For zone 1, certainly, the lower boundary point of zone 1 (X_{CHR,LBZ_1}) denoted the minimum value of CHR as $X_{CHR,LBZ\_1} = 0.215$ in the outer loop.

A middle loop was built to identify the upper boundary point of zone 1 (X_{CHR,UBZ_1}). The initial zone index (k) was set at 1, and $X_{CHR,LBZ\_1} = X_{CHR,MIN} = 0.215$. The inner loop iterates until the R^2 value of the ML predictive sub-model was less than 0.99, which indicates that the ML model does not fit well with the inclusion of more data in this zone. Then the inner loop stopped and returned to the middle loop. X_{CHR,UBZ_1}, which is also the lower boundary point of zone 2 (X_{CHR,LBZ_2}), was identified as $X_{CHR,UBZ\_1} = 0.37$.

For zone 2, a similar procedure was conducted to repeat the middle and inner loops. X_{CHR,UBZ_2} was identified as $X_{CHR,UBZ\_2} = 0.504$. It was found out that a two-zone model was sufficient, and no more zone division was needed. In addition, using a similar approach, it was found out that a one-zone model is sufficient for Opacity prediction.

4.5. Performance evaluation of the zone-based random forest and Catboost models

Seven features, NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V, were selected as the input features to train the zone-based ML models for CE and Opacity. 90% of data were randomly chosen as the training data and the remaining 10% of data were used to test the prediction model, respectively. The zone-based ML CE and Opacity models using RF and Catboost algorithms, respectively, were built in Python. No data from the original experiment datasets was deleted. The prediction accuracy of the predictive sub-models were evaluated based on R^2 and mean squared error (MSE, %). Note that it's easier to use MSE, a differentiable function, to perform mathematical operations, in comparison to a non-differentiable function like MAE. In other word, MAE is more resilient when handling data with many outliers.

5. Results and discussions

5.1. Two-zone RF-CE models

For CE prediction, the original 280 data were divided into two zones based on the CHR value of the vent gas. CHR values represent the compositions of the vent gas. A higher CHR value indicates higher concentrations of longer-chain alkanes or alkenes, as shown in Table 2. For zone 1 ($0.215 \leq CHR \leq 0.370$) and zone 2 ($0.370 < CHR \leq 0.504$), the corresponding zone-based CE predictive models using random forest algorithm, RF-CE $_{zone\ 1}$ and RF-CE $_{zone\ 2}$, were built, respectively. After training the two-zone RF-CE models, the predicted CE values vs. its corresponding test data were presented in Fig. 8 (a). In comparison, an one-zone RF-CE predictive model was trained using all the data to predict CE values, and the predicted CE values vs. its corresponding test data were presented in Fig. 8 (b). Fig. 8 (a) showed that the RF-CE $_{zone\ 1}$ (∗ symbols) and RF-CE $_{zone\ 2}$ (□ symbols) worked extremely well. Clearly, the zone-based RF-CE model outperformed the original one-zone RF-CE model.

Table 2
Examples of the vent gas compositions and the corresponding CHR values.

Compounds	Molecular formula	CHR value
Methane	CH_4	0.250
Ethane	C_2H_6	0.333
Ethylene	C_2H_4	0.500
Propane	C_3H_8	0.375
Propylene	C_3H_6	0.500
Acetylene	C_2H_2	1.000
1,4 Butadiene	C_4H_6	0.667

In Fig. 9, the two-zone-based RF-CE predictive models were depicted using the long dash line for zone 1 ($0.215 \leq CHR \leq 0.37$) and the square dot line for zone 2 ($0.37 < CHR \leq 0.504$), respectively. The R^2 values of these models were 0.9910 and 0.9818 each, and the corresponding MSE values were 0.5624 and 6.4377, respectively. Nevertheless, the R^2 of the one-zone RF-CE model was only 0.8987.

Clearly, the zone-based RF-CE models were more reasonable and accurate than the one-zone RF-CE model. Additionally, the zone-based RF-CE models successfully avoided the scenarios in which a large number of "outliers" have to be omitted for the purpose of increasing the accuracy of a narrowly applicable model.

5.2. One-zone RF-Opacity model

The same seven features, i.e., NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V, were used as the input features to train the opacity model. The one-zone RF-Opacity model was built, fully utilizing 280 datasets in total, in which 252 datasets were randomly selected as training the data while the subset of the remaining 28 datasets were used for testing the model. The R^2 value was calculated to evaluate the performance of the RF-Opacity model. Fig. 10 compares the predicted value of Opacity vs. the test data, where the symbol "o" denotes the RF-Opacity prediction results and the symbol "Δ" denotes the experimental data. The RF-Opacity model achieved the R^2 value of 0.992 and the corresponding MSE value of 0.231.

5.3. Comparison of prior models and zone-based ML models

Compared with the validation of the prior flare performance models, general quadratic response surface models, ANN algorithm models, and RF algorithm models, for CE prediction, the zone-based models achieve high predictive accuracy ($R^2 \geq 0.98$) without excluding any original data; its prediction accuracy is higher than GQM's 0.920 with 77 outliers, ANNs model's 0.929 (R value), and the same as prior RF model's 0.992 with 55 outliers. This comparison is listed in Table 3. For Opacity prediction, the zone-based models achieve high predictive accuracy ($R^2 \geq 0.99$) without deleting outliers; its accuracy is better than GQM's 0.900 with 96 outliers, ANNs model's 0.949 (R value), and the same as prior RF model's 0.936 with 55 outliers.

The zone-based CE models gives better results because carbon and hydrogen atomic ratio (CHR) values reflect the compositions of the vent gas, which affects the combustion mechanism and the eventual combustion efficiency significantly. A higher CHR value indicates higher concentrations of longer-chain alkanes, alkenes or alkynes, as shown in Table 2. The algorithm used CHR values to divide the entire dataset into two zones, zone 1

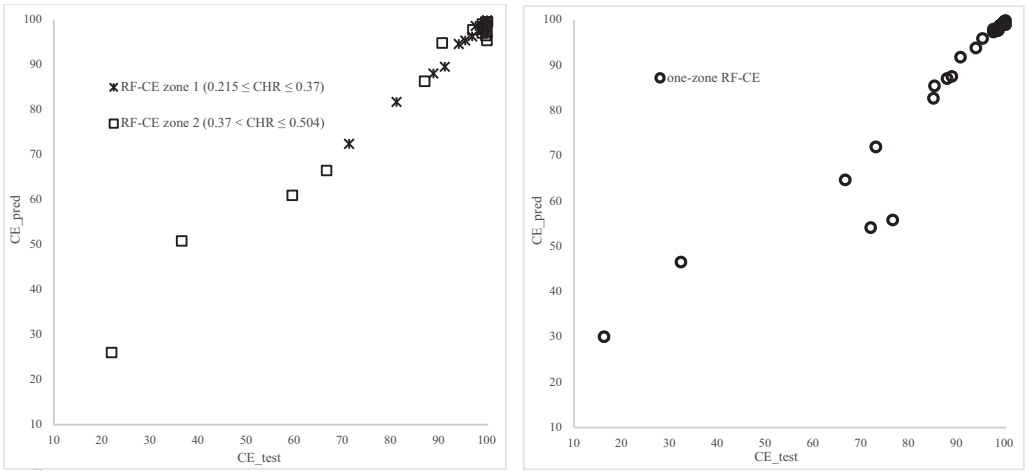

Fig. 8. Predicted CE vs. experimental CE.
(a) The zone-based RF models (b) The one-zone RF model.

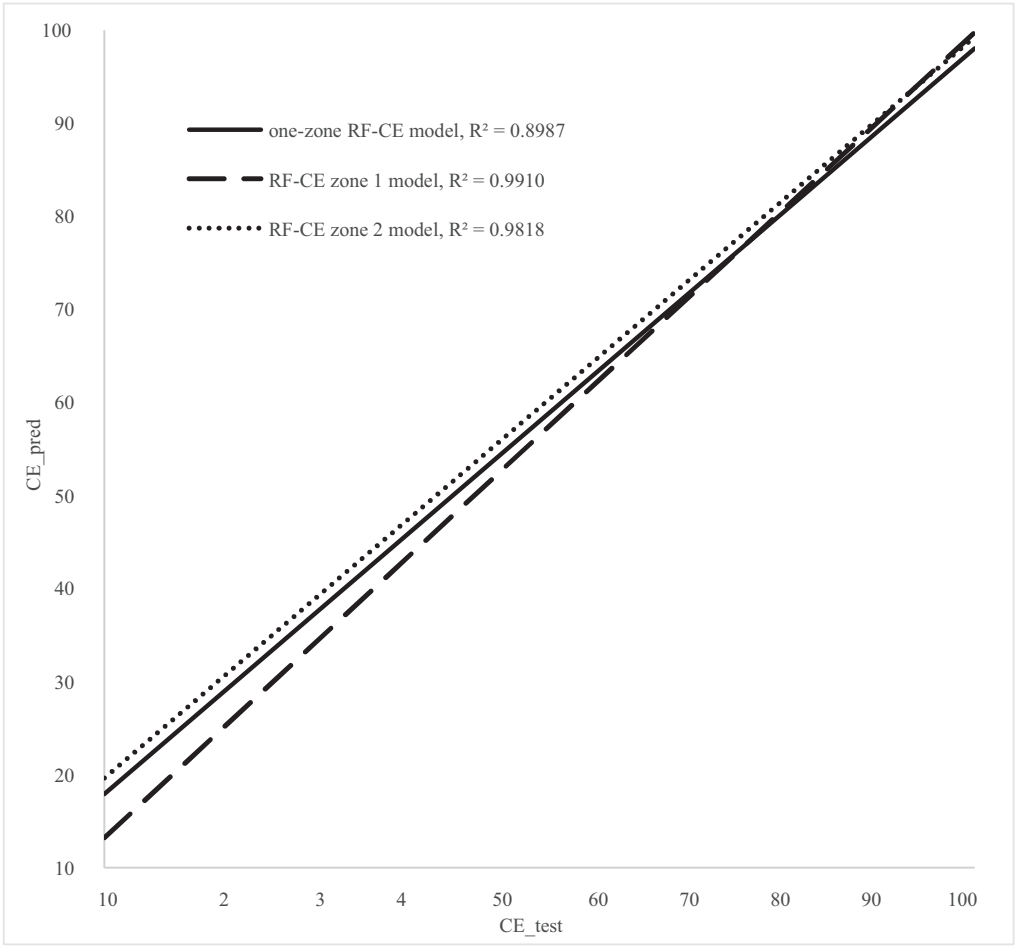

Fig. 9. The trendlines of predicted CE vs. experimental CE.

(0.215 ≤ CHR ≤ 0.370) and zone 2 (0.370 < CHR ≤ 0.504). The CHR value of different molecules are shown in Table 2. This indicates datasets in zone 1 representing vent gas with more alkanes while datasets in zone 2 representing vent gas with more longer-chain alkanes, alkenes and/or alkynes. Due to the impact of the double bound(s) in alkene and triple bound(s) in alkyne, their combustion mechanisms are different from that of alkanes. While for Opacity prediction, a one-piece ML model is sufficient.

Overall, the zone-based models for CE and Opacity predictions have significantly higher predictive accuracies than those of the prior three one-piece models and none of the original data were deleted.

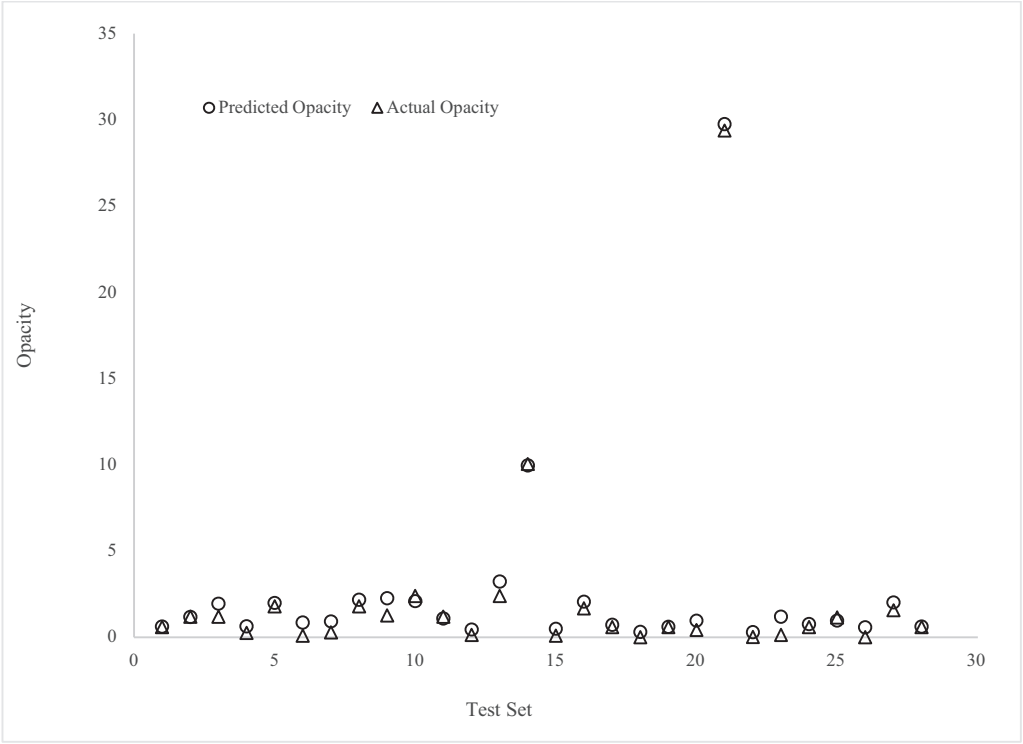

Fig. 10. Predicted opacity vs. experimental opacity.

Table 3
Summary of different flare performance of ML models.

Model	Output	Features	Outliers	R^2 / R	MAE / MSE
GQM Alphones et al. (2020)	Logit (100 - %CE)	CHR, CN, NHV_{cz}, U/V, S, and U	77	0.920	2.100
	Logit (%Opacity)	CHR, CN, NHV_{cz}, V, S, and U	96	0.900	0.940
ANNs Damodara et al. (2020)	Logit (100 - %CE)	CHR, CN, NHV_{cz}, U/V, S, U, and D	-	0.929 (R)	1.390
	Logit (%Opacity)	CHR, NHV_{cz}, and U/V	-	0.949 (R)	1.280
RF Wang (2019)	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	55	0.992	1.320
	Opacity	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	55	0.936	0.980
One-zone RF	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.899	4.562 (MSE)
Zone-based RF (Zone 1)	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.991	0.562 (MSE)
Zone-based RF (Zone 2)	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.982	6.438 (MSE)
One-zone RF	Opacity	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.992	0.231 (MSE)
Zone-based Catboost (Zone 1)	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.998	0.141 (MSE)
Zone-based Catboost (Zone 2)	CE	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.991	3.256 (MSE)
One-zone Catboost	Opacity	NHV_{vg}, NHV_{cz}, S, CHR, D, U, and U/V	0	0.996	1.287 (MSE)

5.4. Comparison of the zone-based Catboost and RF algorithms models

For CE prediction using the Catboost algorithm in zone 1 ($0.215 \leq CHR \leq 0.370$) and zone 2 ($0.370 < CHR \leq 0.504$), the R^2 values of these models were 0.998 and 0.982 each, and the corresponding MSE values were 0.141 and 3.256, respectively. For a one-zone Catboost-Opacity model, the R^2 value of 0.996 and the corresponding MSE value of 1.287. Clearly, the zone-based Catboost models were as accurate as the zone-based RF models. This shows zone partition is the most important step in model development.

6. Conclusions

This study developed a novel zone-based ML approach that is applicable when the modeler does not know which feature is the best one for data partition. This algorithm can figure out the best data partition approach automatically and develop high-performance ML models. This algorithm was applied to the prediction of the flaring performance. The novel zone-based ML approach eventually found out the carbon and hydrogen atomic ratio (CHR) is the best choice for data partition when the modeler tries to predict combustion efficiency. While for Opacity prediction, a one-piece ML model is sufficient.

The zone-based models for CE and Opacity predictions have significantly higher predictive accuracies than those of the prior one-piece models, and none of the original data were deleted. In addition, both the zone-based RF and Catboost models for CE and Opacity achieved superior predictive performance, and no original data were deleted.

It was also found that the appropriate zone partition is the most important factor in the modeling of complicated processes. These high prediction accuracies achieved by zone-based models are notable, particularly with regard to the models' simplicity, general applicability, and high reliability in the broader fields of engineering.

Theoretically, the zone-based ML approach is also applicable when multiple features are selected for zone partition. The modeler needs to try different combinations of the features to find out which one is the best. It will take much more computational time

than exploring every single feature. In this case, domain knowledge can help the modeler narrow down the list of combinations and reduce the computational workload.

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.

CRediT authorship contribution statement

Helen H. Lou: Supervision, Methodology, Conceptualization, Writing – review & editing, Project administration. **Jian Fang:** Methodology, Visualization, Writing – original draft, Writing – review & editing. **Huilong Gai:** Validation, Visualization, Writing – review & editing. **Richard Xu:** Visualization, Writing – review & editing. **Sidney Lin:** Supervision, Conceptualization, Writing – review & editing.

Acknowledgment

This research was supported by Lamar University Center for Midstream Management and Science (CMMS) and Lamar University College of Engineering. The authors also wish to thank Dr. Sujing Wang for helpful conversations and supports.

References

Abubakirov, R., Yang, M., Khakzad, N., 2020. A risk-based approach to determination of optimal inspection intervals for buried oil pipelines. Process Saf. Environ. Prot. 134, 95–107. doi:10.1016/j.psep.2019.11.031.

Albright, T.A., Burdett, J.K., Whangbo, M.H., 2013. Orbital Interactions in Chemistry, (2nd ed.) John Wiley & Sons, Inc Second Edi. ed doi:10.1002/9781118558409.

Allen, D.T., Torres, V.M., 2011. TCEQ 2010 flare study final report. Austin, TX.

Allen, D.T., Torres, V.M., 2011. TCEQ 2010 flare study final report - appendices. Austin, TX.

Alphones, A., Damodara, V., Wang, A., Lou, H., Li, X., Martin, C.B., Chen, D.H., Johnson, M.R., 2020. Response surface modeling and setpoint determination of steam- and air-assisted flares. Environ. Eng. Sci. 37, 246–262. doi:10.1089/ees.2019.0089.

Hughes, B., 2019. flare.IQ. Baker Hughes Company.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. Noise Reduct. Speech Process 2, 37–40. doi:10.1007/978-3-642-00296-0_7.

Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197–227. doi:10.1007/s11749-016-0481-7.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1201/9780429469275-8.

Cade, R., Evans, S., 2010. Performance test of a steam-assisted elevated flare with passive FTIR - Marathon Petroleum Company, LLC, Texas City, Texas. Clean Air Engineering, Inc.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R.D.P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Comput. Ind. Eng. 137, 1–10. doi:10.1016/j.cie.2019.106024.

Çatak, F.Ö., 2017. Classification with boosting of extreme learning machine over arbitrarily partitioned data. Soft Comput. 21, 2269–2281. doi:10.1007/s00500-015-1938-4.

Chan, P.K., Stolfo, S.J., 1995. A comparative evaluation of voting and meta-learning on partitioned data. In: Proceedings of the Twelfth International Conference on Machine Learning. Tahoe City, California. Morgan Kaufmann Publishers, Inc. doi:10.1016/B978-1-55860-377-6.50020-7.

Chrysafis, I., Mallinis, G., Gitas, I., Tsakiri-Strati, M., 2017. Estimating Mediterranean forest parameters using multi seasonal Landsat 8 OLI imagery and an ensemble learning method. Remote Sens. Environ. 199, 154–166. doi:10.1016/j.rse.2017.07.018.

Corbin, D.J., Johnson, M.R., 2014. Detailed expressions and methodologies for measuring flare combustion efficiency, species emission rates, and associated uncertainties. Ind. Eng. Chem. Res. 53, 19359–19369. doi:10.1021/ie502914k.

Damodara, V.D., Alphones, A., Chen, D.H., Lou, H.H., Martin, C., Li, X., 2020. Flare performance modeling and set point determination using artificial neural networks. Int. J. Energy Environ. Eng. 11, 91–109. doi:10.1007/s40095-019-00314-3.

Ewing, B., Roesler, D., Evans, S., 2010. Performance test of a steam-assisted elevated flare with passive FTIR - Marathon Petroleum Company, Detroit, MI. Clean Air Engineering, Inc.

Gomez, C., Mangeas, M., Petit, M., Corbane, C., Hamon, P., Hamon, S., De Kochko, A., Le Pierres, D., Poncet, V., Despinoy, M., 2010. Use of high-resolution satellite imagery in an integrated model to predict the distribution of shade coffee tree hybrid zones. Remote Sens. Environ. 114, 2731–2744. doi:10.1016/j.rse.2010.06.007.

Hancock, J.T., Khoshgoftaar, T.M., 2020. Catboost for big data: an interdisciplinary review. J. Big Data 7. doi:10.1186/s40537-020-00369-8.

He, G., Zhou, C., Luo, T., Zhou, L., Dai, Y., Dang, Y., Ji, X., 2021. Online optimization of fluid catalytic cracking process via a hybrid model based on simplified structure-oriented lumping and case-based reasoning. Ind. Eng. Chem. Res. 60, 412–424. doi:10.1021/acs.iecr.0c04109.

Jäger, C., Henning, T., Schlögl, R., Spillecke, O., 1999. Spectral properties of carbon black. J. Non Cryst. Solids 258, 161–179. doi:10.1016/S0022-3093(99)00436-6.

Johnson, M., 2014. Flare efficiency & emissions : past and current research. Global Forum on Flaring and Venting Reduction and Natural Gas Utilization. Carleton University, Ottawa, ON, Canada.

Li, B., Gong, A., Zeng, T., Bao, W., Xu, C., Huang, Z., 2022. A zoning earthquake casualty prediction model based on machine learning. Remote Sens. 14, 1–27. doi:10.3390/rs14010030.

Lou, H.H., Gai, H., 2020. How AI can better serve the chemical process industry. In: Hydrocarbon Processing Special Focus: the Digital Plant. Hydrocarbon Processing Magazine, pp. 37–39.

Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., Sun, Y., 2021. Combination of feature selection and Catboost for prediction: the first application to the estimation of aboveground biomass. Forests 12, 1–22. doi:10.3390/f12020216.

Maddala, G.S., 1992. Introduction to Econometrics, (3rd ed.) John Wiley & Sons, Ltd Chichester, New York.

Louppe, G., 2014. Understanding random forests: from theory to practice. PhD dissertation, Department of Electrical Engineering & Computer Science, University of Liège, doi:10.13140/2.1.1570.5928.

Mohammadi, M., Amir, S., Motevalli, A., Hashemi, H., 2022. Human-induced arsenic pollution modeling in surface waters - an integrated approach using machine learning algorithms and environmental factors. J. Environ. Manag. 305, 114347. doi:10.1016/j.jenvman.2021.114347.

Nguyen, H., Bui, X.N., Choi, Y., Lee, C.W., Armaghani, D.J., 2021. A novel combination of whale optimization algorithm and support vector machine with different kernel functions for prediction of blasting-induced fly-rock in quarry mines. Nat. Resour. Res. 30, 191–207. doi:10.1007/s11053-020-09710-7.

McDaniel, M., Tichenor, B.A., 1983. Flare efficiency study. US Environmental Protection Agency. URL https://www.tceq.texas.gov/assets/public/implementation/air/rules/Flare/Resource_1.pdf (accessed 12.2.16).

Pohl, J.H., Payne, R., Lee, J., 1984. Evaluation of the efficiency of industrial flares: test results. Final report Oct 80-Feb 84. URL https://nepis.epa.gov/Exe/ZyNET.exe/P100RIYK.TXT?ZyActionD=ZyDocument&Client=EPA&Index=1981+Thru+1985&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=&File=D%3A%5Czyfiles%5CIndex%20Data%5C81thru85%5CTxt%5C00000028%5CP100RIYK.txt&User=ANONYMOUS&Password=anonymous&SortMethod=h%7C-&MaximumDocuments=1&FuzzyDegree=0&ImageQuality=r75g8/r75g8/x150y150g16/i425&Display=hpfr&DefSeekPage=x&SearchBack=ZyActionL&Back=ZyActionS&BackDesc=Results%20page&MaximumPages=1&ZyEntry=1&SeekPage=x&ZyPURL.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2019. Catboost : unbiased boosting with categorical features. Mechaine Learning.

Qin, S.J., Chiang, L.H., 2019. Advances and opportunities in machine learning for process data analytics. Comput. Chem. Eng. 126, 465–473. doi:10.1016/j.compchemeng.2019.04.003.

Scornet, E., 2015. Random Forests and Kernel Methods. Cornell University doi:10.1109/TIT.2016.2514489.

Singh, K.D., Gangadharan, P., Chen, D.H., Lou, H.H., Li, X., Richmond, P., 2014. Computational fluid dynamics modeling of laboratory flames and an industrial flare. J. Air Waste Manag. Assoc. 64, 1328–1340. doi:10.1080/10962247.2014.948229.

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. doi:10.1016/j.apenergy.2018.02.126.

Trivanovic, U., Sipkens, T.A., Kazemimanesh, M., Baldelli, A., Jefferson, A.M., Conrad, B.M., Johnson, M.R., Corbin, J.C., Olfert, J.S., Rogak, S.N., 2020. Morphology and size of soot from gas flares as a function of fuel and water addition. Fuel 279. doi:10.1016/j.fuel.2020.118478.

US EPA, 2016. 40 CFR § 63.670 - Requirements for flare control devices. Code Fed. Regul. URL https://www.law.cornell.edu/cfr/text/40/63.670 (accessed 3.26.20).

US EPA, 2006. AP 42, Fifth Edition, Volume I Chapter 13: Miscellaneous Sources. US EPA. URL https://www3.epa.gov/ttnchie1/ap42/ch13/ (accessed 11.10.20).

Wang, A., 2019. Combustion Kinetics Study and Machine Learning for the Simulation and Control of Industrial Flares. Lamar University.

Wang, R., Bao, J., Yao, Y., 2019. A data-centric predictive control approach for nonlinear chemical processes. Chem. Eng. Res. Des. 142, 154–164. doi:10.1016/j.cherd.2018.12.002.

Ye, H., Liu, H., 2005. Approach to modelling feature variability and dependencies in software product lines. IEE Proc. Softw. 152, 101–109. doi:10.1049/ip-sen:20045007.

Zeng, Y., Morris, J., Dombrowski, M., 2016. Validation of a new method for measuring and continuously monitoring the efficiency of industrial flares. J. Air Waste Manag. Assoc. 66, 76–86. doi:10.1080/10962247.2015.1114045.

Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L., Subramanian, R., 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech. 11, 291–313. doi:10.5194/amt-11-291-2018.