ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

http://pubs.acs.org/journal/acsodf

Article

Feasibility of the Optimal Design of AI-Based Models Integrated with Ensemble Machine Learning Paradigms for Modeling the Yields of Light Olefins in Crude-to-Chemical Conversions

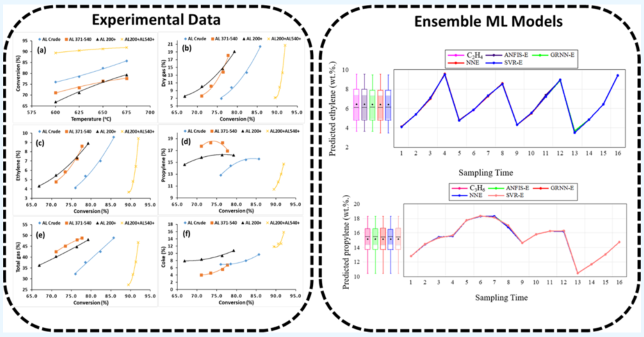
A. G. Usman, Abdulkadir Tanimu,* S. I. Abba, Selin Isik, Abdullah Aitani, and Hassan Alasiri

Cite This: *ACS Omega* 2023, 8, 40517−40531

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The prediction of the yields of light olefins in the direct conversion of crude oil to chemicals requires the development of a robust model that represents the crude-to-chemical conversion processes. This study utilizes artificial intelligence (AI) and machine learning algorithms to develop single and ensemble learning models that predict the yields of ethylene and propylene. Four single-model AI techniques and four ensemble paradigms were developed using experimental data derived from the catalytic cracking experiments of various crude oil fractions in the advanced catalyst evaluation reactor unit. The temperature, feed type, feed conversion, total gas, dry gas, and coke were used as independent variables. Correlation matrix analyses were conducted to filter the input combinations into three different classes (M1, M2, and M3) based on the relationship between dependent and independent variables, and three performance metrics comprising the coefficient of determination (R^2), Pearson correlation coefficient (PCC), and mean square error (MSE) were used to evaluate the prediction performance of the developed models in both calibration and validations stages. All four single models have very low R^2 and PCC values (as low as 0.07) and very high MSE values (up to 4.92 wt %) for M1 and M2 in both calibration and validation phases. However, the ensemble ML models show R^2 and PCC values of 0.99−1 and an MSE value of 0.01 wt % for M1, M2, and M3 input combinations. Therefore, ensemble paradigms improve the performance accuracy of single models by up to 58 and 62% in the calibration and validation phases, respectively. The ensemble paradigms predict with high accuracy the yield of ethylene and propylene in the catalytic cracking of crude oil and its fractions.

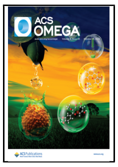


1. INTRODUCTION

Demand for petrochemical feedstocks has been steadily increasing due to the growing demand for petrochemical intermediates that serve as raw materials in the production chain of many industries.[1] Light olefins, particularly ethylene and propylene, are an important part of these petrochemical feedstocks that have seen increasing demand in recent times.[2] The global light olefin market is projected to register a compound annual growth rate of about 5.6% by the year 2030.[3] Therefore, there is an urgent need to expand the production capacity for these olefins. At present, steam cracking of naphtha is the main source of light olefins, covering about 80% of all of the light olefin sources. Because of the large energy consumption, naphtha price dynamics, and large CO_2 generation during steam cracking of naphtha, giant refineries are considering direct production of light olefins from relatively cheap feed, the crude oil, via catalytic cracking.[4] The future refinery targets optimizing the refinery process to allow simultaneous production of transportation fuel and light olefins.[5]

Research in catalytic cracking of crude oil to light olefins has witnessed great improvement in terms of refining technology, process intensification, process optimization, and catalyst formulations in the last few decades.[6−8] For example, the downer reactor technology was developed to minimize the feed residence time, preventing overcracking, hydrogen transfer reactions, and aromatization.[9] Alabdullah et al.[10] proposed a multizone fluidized bed catalytic reactor that performed in situ catalyst stripping and regeneration, thus literally performing all of the refining steps in a single step. The role of catalyst formulation in modulating the light olefin selectivity in crude oil catalytic cracking, such as the addition of ZSM-5 additives, and tuning of zeolite porosity, zeolite Si/Al ratio, and matrix effect

Received: July 20, 2023
Accepted: October 9, 2023
Published: October 18, 2023



has been reviewed by Tanimu et al.[5] Additionally, process optimization of crude oil catalytic cracking by varying reactor temperatures in the range of 475−550 °C and catalyst-to-oil ratios (C/O) in the range of 1−4 g/g was performed in a microactivity unit (MAT), and it was discovered that 550 °C and 3 g/g are the optimum reactor temperature and C/O, respectively.[11] Similarly, the reactor temperature of an advanced catalytic evaluation (ACE) unit was varied between 550 and 650 °C during the catalytic cracking of Arab Super Light (ASL) crude oil, and it was discovered that conversion and product yield (including dry gas and coke) increased steadily.[12] The endothermicity of hydrocarbon cracking is typically responsible for the observed increase in the conversion and product yield with temperature. Although high temperatures result in high conversion and light olefin selectivity, the formation of large amounts of dry gas (H_2 and C_1−C_2) and coke under a high-temperature cracking process is not desirable in refinery processes.[13] High temperatures are linked to monomolecular pyrolytic cracking, where a carbonium ion collapses and preferentially forms H_2, CH_4, C_2H_6, and a carbenium ion that later transforms to alkene.[14] Thus, increasing the reactor temperature for high light olefin yields comes with an expensive choice of producing more dry gas and coke. However, with careful optimization of reactor temperature, an optimum might be reached where the dry gas and coke yield is brought to a minimum while ensuring maximum yield of light olefin. Additionally, the type of crude oil feed has been related to the light olefins and dry gas and coke yields in previous studies. Comparing the catalytic cracking of ASL, AXL, and AL crude oil feeds over the E-Cat/Z80 catalyst, Usman et al.[11] discovered that the total light olefin (C_2H_4−C_4H_8) yield is highest in the cracking of AXL crude oil feed, while the coke and dry gas yield is lowest in the ASL crude oil feed. Therefore, with careful optimization of the crude oil feed and cracking temperature, the yield of light olefins can be maximized while lessening the yield of dry gas and coke. However, this approach involves handling a large number of variables that are ordinarily difficult to process using simplified/physical models. The re-emergence of artificial intelligence (AI) in the early 21st century has made the analysis and optimization of a large number of variables in various academic and industrial application much easier and less laborious.[15] For instance, a comparative study between the artificial neural network (ANN) model and the nonlinear statistical model showed that the ANN model has higher prediction accuracy.[16] A hybrid neural network model with a physical reaction model was discovered to be more efficient in yield optimization and prediction than a conventional reaction model.[17] Recently, Kawai et al.[18] developed an AI hybrid reaction model for the optimization of the catalyst makeup rate in a residue fluid catalytic cracking process, which maximized the light olefin yield, minimized the catalyst loss, and offered yield prediction with a high level of accuracy.

Even though there is a recent establishment of various modeling processes in the field of catalytic cracking, the efficiency of some chemometrics processes such as AI and machine learning (ML)-based approaches is associated with numerous limitations. Nevertheless, to improve their accuracies, a new novel technique of ensemble learning (EL) depicts reliable performance in various fields of chemometrics and cheminformatics.[19−21] EL has the ability to capture the limitations that can be depicted by the best standalone ML-based chemometric model and hence has the ability to improve the performance efficiency. Additionally, the catalytic cracking process modeling is attributed to different chemical and physical parameters that describe it as intricate for proper feature selection. Proper input combination selection has been utilized in order to understand the input−output relation. The basic motivation of the current work is the introduction of the EL approach for modeling the catalytic cracking yield of ethylene and propylene hydrocarbons. Moreover, the current research aimed at modeling the catalytic cracking yield using standalone techniques, namely, regression tree (RT), least-squares boosting (LSQ-BOOST), Gaussian process regression (GPR), and robust linear regression (RLR) models. Afterward, the simulated values were improved using four ensemble paradigms, namely, generalized regression neural network ensemble (GRNN-E), support vector regression ensemble (SVR-E), feedforward neural network ensemble (NNE), and adaptive neuro-fuzzy inference system for modeling the catalytic cracking yield of ethylene and propylene. To the best of the authors' knowledge, this is the first work in the technical literature that depicts both the implementation of correlational-based feature selection and nonlinear EL techniques for modeling the yield of ethylene and propylene in crude oil catalytic cracking reactions.

2. EXPERIMENTAL SECTION

2.1. Catalytic Cracking Reaction. The catalytic cracking reaction was carried out according to the ASTM D-7964 method in an advanced catalyst evaluation (ACE) unit. The unit is characterized by an automated fixed-fluidized bed reactor and was manufactured by Kayser Technology Inc. Four kinds of feeds were cracked using a steamed deactivated commercial catalyst at a constant CTO ratio of 5. Thus, a catalyst weight of 7.5 g was loaded into a catalyst hopper, and a feed weight of 1.5 g was injected into the reactor at a rate of 2 g/min, resulting in a time on stream of 45 s. The reactor temperature was varied between 600 and 675 °C while maintaining atmospheric pressure conditions.

After the reaction, N_2 gas was purged into the reactor for about 9 min for catalyst stripping. The liquid product was collected in a glass receiver that was fixed to the reactor exit and kept in a water bath at a temperature of −1.25 °C. The total gaseous product weight was measured by water displacement in a gas receiver. After catalyst stripping was finished, the catalyst was regenerated by switching the N_2 valve with air and increasing the reactor temperature to 700 °C. Thus, all of the coke deposited on the catalyst was oxidized to CO_2 and subsequently analyzed by online IR spectroscopy.

2.2. Analysis ofCracked Products. The gaseous products were analyzed online using an Agilent 3000A micro gas chromatograph equipped with a multicolumn system, a multichannel system, and four thermal conductivity detectors (TCDs). The liquid products were analyzed based on the boiling point distribution using a Shimadzu GC 2010 Plus, Japan, with a flame ionization detector (FID) and according to the ASTM D-2887 for standard simulated distillation (SimDist). This boiling point distribution resulted in three liquid fractions defined as gasoline (C5−221 °C), light cycle oil (LCO, 221−343 °C), and heavy cycle oil (HCO, 343+ °C). The results of gas and liquid product analyses were used to calculate conversion and product yields for the three different feeds. The conversion is calculated as follows

$$\%\ \text{Conversion} = (100 - \text{LCO} - \text{HCO}) \quad (1)$$

2.3. Proposed Chemometric-Based Methodology. The data set utilized in the current study was generated from the

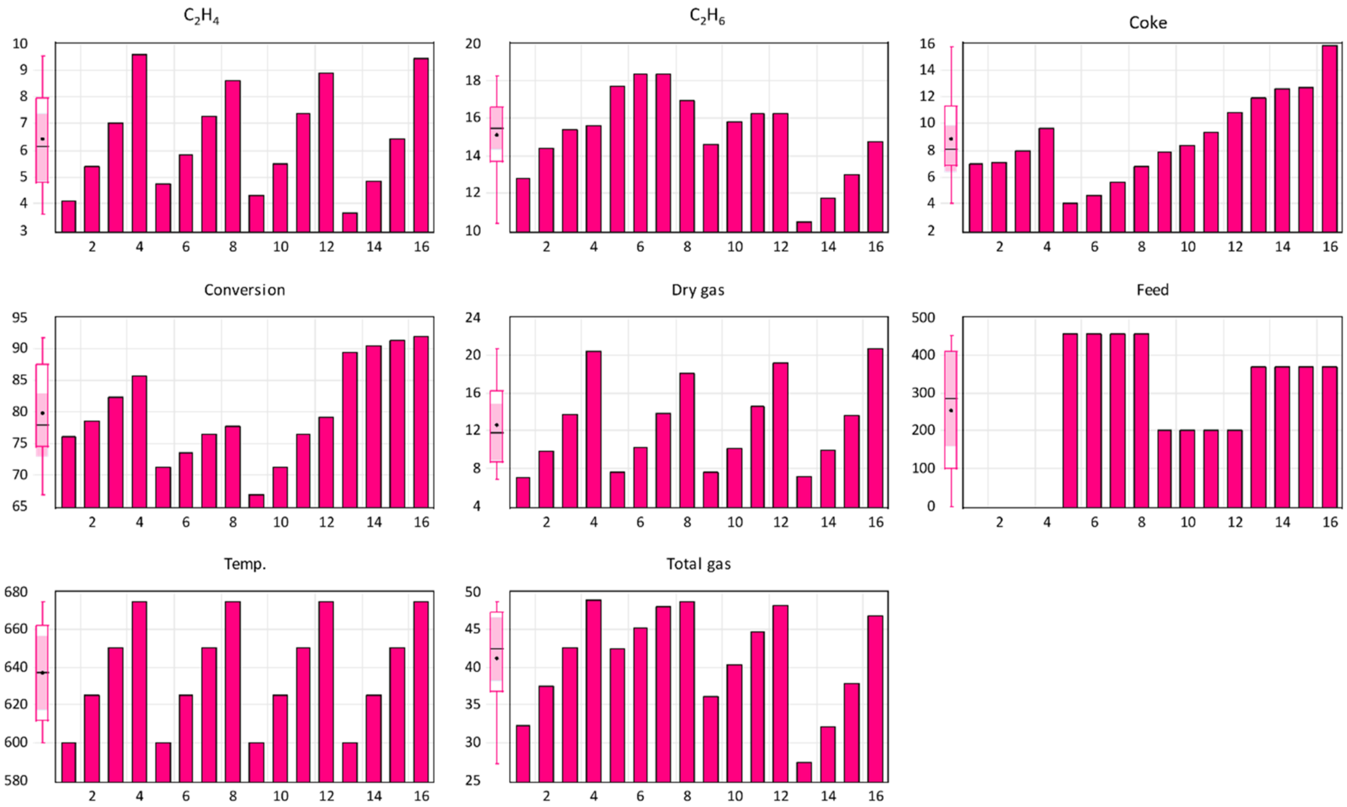


Figure 1. Data distribution plot embedded with the bar−box plot graph of the employed independent−dependent variables.

ACE experiments, consisting of both physical and chemical features as the input and output variables. This study consists of temperature, feed type, feed conversion, total gas, dry gas, and coke as independent variables. The catalytic cracking yields of ethylene and propylene were two dependent variables. Moreover, chemometric-based data analysis by AI and ML depends heavily on data preprocessing that entails a suitable approach to formatting the data sets into a suitable format as well as to understanding the physical and chemical interaction of the variables using techniques such as sensitivity analysis and feature selection. In this study, Microsoft Excel 2019 was used to perform preliminary data cleaning in order to exclude any mistakes, inconsistencies, and potentially misleading information. The Microsoft Excel in-built data cleaning capabilities were used in eliminating duplicates, fixing mistakes, and adding missing information. Second, the noisy and irrelevant data and outliers, which could skew the result of the study, were also removed. All variables in the data set (Figure 1) were standardized to ensure that they were of the same size and range. This ensured that bias was eliminated and the analytical performance was improved. Subsequently, the modeling was done using MATLAB (2020a). Furthermore, Figure 1 illustrates the distribution of the data, which refers to the way data points are spread or organized within a data set. The current study involves 16 instances that are spread differently by each variable used during the simulation process. Also, it describes how frequently different values or ranges of values appear in the data set. Understanding the data distribution is essential in statistics and data analysis because it can impact the choice of statistical methods, modeling techniques, and interpretation of results. Therefore, four standalone models, including three nonlinear techniques: RT, LSQ-BOOST, and GPR, as well as the classical linear RLR model, were simulated on the regression learner from MATLAB (2020a). Hence, four different EL techniques indicating GRNN-E, SVR-E, NNE, and ANFIS-E were used in improving the performance of the standalone models.

2.4. Regression Tree (RT).

A regression tree (RT) is an ML algorithm that builds a decision tree to predict a numerical value (i.e., a continuous output variable) based on a set of input features. At a high level, RT recursively partitions the feature space into regions or leaves, each with its own predicted output value. The partitioning is performed by selecting a feature and a threshold that best splits the data into two subsets based on the output variable.[22] The splitting process is continued until a stopping criterion is reached such as a maximum tree depth or a minimum number of samples per leaf. Furthermore, to make a prediction for a new data point, the RT navigates down the tree from the root node to a leaf node based on the values of the input features and returns the predicted output value associated with the leaf node.[23]

RTs have several advantages, such as being able to handle nonlinear relationships between the input features and the output variable, being interpretable (as the decision rules can be visualized), and being robust to outliers and missing data. However, they may also suffer from overfitting if the tree is too deep or if the number of features is large compared to the number of samples.[24]

2.5. Gaussian Process Regression (GPR).

Gaussian process regression (GPR) is considered to be a strong and robust AI-based model, nonparametric, probabilistic, supervised, and unsupervised learning approach, which generalizes complex and nonlinear function mapping.[25] GPR has recently gained more attention from various modelers and forecasters from different fields ranging from medical sciences to engineering and technology. This is due to the fact that GPR has the ability to handle highly nonlinear phenomena owing to the implementation of Kernel functions. Furthermore, this model equally has the ability to provide reliable responses to the input data.[26]

2.6. Least-Squares Boosting (LSQ-BOOST). Least-squares boosting (LSQ-BOOST) is an ML algorithm that combines the power of boosting and least-squares regression. It is used to solve regression problems, where the goal is to predict a continuous output variable.[27] The LSQ-BOOST algorithm works by iteratively fitting a regression model to the data and then adjusting the weights of the training samples to emphasize the points that the model is currently not able to predict well. This approach is known as boosting and is a powerful technique for improving the accuracy of weak learners. In LSQ-BOOST, the regression model is typically a simple linear regression or a decision tree regression, which is trained on a subset of the data with current weights assigned to each sample. The weights are updated using the least-squares algorithm, which minimizes the sum of squared errors between the predicted and actual output values.[28] During each iteration of the algorithm, the weights of the misclassified samples are increased, while the weights of the correctly classified samples are decreased. This results in a sequence of models, each of which focuses on the difficult-to-predict samples that were missed by the previous models.[28]

2.7. Robust Linear Regression (RLR). Robust linear regression (RLR) is a statistical technique that is used to model the relationship between a dependent variable and one or more independent variables. The main difference between RLR and standard linear regression is that RLR is less sensitive to outliers in the data.[29] Standard linear regression assumes that the errors in the data are normally distributed and have a constant variance. However, when there are outliers in the data, the assumptions of standard linear regression are often violated, and the resulting model can be biased and inefficient.[30] Standard linear regression assumes that the errors in the data are normally distributed and have constant variance.[31] However, when there are outliers in the data, the assumptions of standard linear regression are often violated, and the resulting model can be biased and inefficient. In RLR, the aim is to find a line that fits the data as well as possible while minimizing the influence of outliers. One popular method for RLR is called the "M-estimator," which is based on minimizing a weighted sum of squared residuals. The weights are chosen to downweight the influence of outliers so that they have less impact on the final model.[32]

Thus, it also examines the interplay between variables and describes their relationship by modifying only one independent variable. Correlations between n regressor factors and dependent variable y are shown in eq 2.

$$y = b_0 + b_1x_1 + b_2x_2 + ...b_ix_i \tag{2}$$

Equation 2 gives a simple representation of the values ith as a predictor. Therefore, b represents the coefficient of the ith predictor, while b_0 represents the constant for regression, with ξ as the error.[33,34] Generally, there are different kinds of linear regressions (LR), including multilinear regression (MLR), stepwise linear regression (SWLR), interaction linear regression (ILR), and robust linear regression (RLR), as can be seen from these studies.[35] In the current study, the ILR method was employed.

2.8. Ensemble Learning Techniques. Ensemble learning is an ML technique that involves combining multiple models to improve their overall performance. There are several ensemble learning techniques. The first one is bagging (bootstrap aggregating), which involves training multiple models independently on randomly sampled subsets of the training data and then combining their predictions by taking the average or majority vote. Bagging is particularly effective for reducing the variance of unstable models, such as decision trees. The second one is boosting, which involves training a sequence of models where each subsequent model focuses on correcting the errors made by the previous model. Boosting algorithms, such as AdaBoost and gradient boosting, can significantly improve the accuracy of weak learners and have been widely used in various applications. The third one is composed of an ensemble of heterogeneous models, which can also combine models from different families, such as combining decision trees with neural networks or support vector machines. This approach can improve the robustness of the ensemble and exploit the strengths of different families of models. The fourth one is stacking EL, which involves training multiple models on the same data and using their predictions as input features to a final meta-model. The meta-model learns to combine the predictions of the base models to make the final prediction. Stacking can improve the predictive performance of the individual models by leveraging their complementary strengths. In the current study, the stacking EL-based approach was utilized by using the results of the standalone models (RT, LSQ-BOOST, GPR, and RLR) as independent variables, while the catalytic cracking yields of ethylene and propylene were maintained as dependent variables. Hence, the following four different ensemble techniques were trained based on this idea: generalized neural network ensemble (GRNN-E), support vector regression ensemble (SVR-E), neural network ensemble (NNE), and adaptive neuro-fuzzy inference system ensemble (ANFIS-E).

2.8.1. Generalized Neural Network Ensemble (GRNN-E). Multiple models, frequently of the same type, are trained using the machine learning technique known as ensemble learning to address a problem. According to the theory, utilizing numerous models in combination rather than just one can result in a better overall performance.

Therefore, for this case, GRNN-E was used in training the outcomes derived from the four different standalone models (RT, LSQ-BOOST, GPR, and RLR), which were considered as input variables. GRNN is a kind of ANN technique that is generally used in regression operations. It is renowned for its capacity to make predictions based on input data and to approximate functions. It contains different overviews and an architecture composed of four layers: the input layer, the pattern layer, the summation layer, and the output layer. Each layer consists of a group of neurons. GRNN training requires labeled input−output pairs and is noniterative. The network records the input patterns and the accompanying outputs during training. The pattern layer determines how comparable the raw data and patterns that have been stored are.

This model is developed based on the idea of regression analysis that follows a nonlinear pattern. Let $f(x, y)$ denote the joint probability density function of vector random variable X and scalar random variable y. The estimated value of Y can be obtained by using the following equation (eq 3).

$$Y = E[y|X] = \int_{-\infty}^{\infty} yf(X, y)\mathrm{d}y / \int_{-\infty}^{\infty} f(X, y)\mathrm{d}y \tag{3}$$

2.8.2. Support Vector Regression Ensemble (SVR-E). SVR is a regression technique that builds on the principles of support vector machines (SVMs), which are commonly used for regression tasks. SVR aims to find a hyperplane that best fits

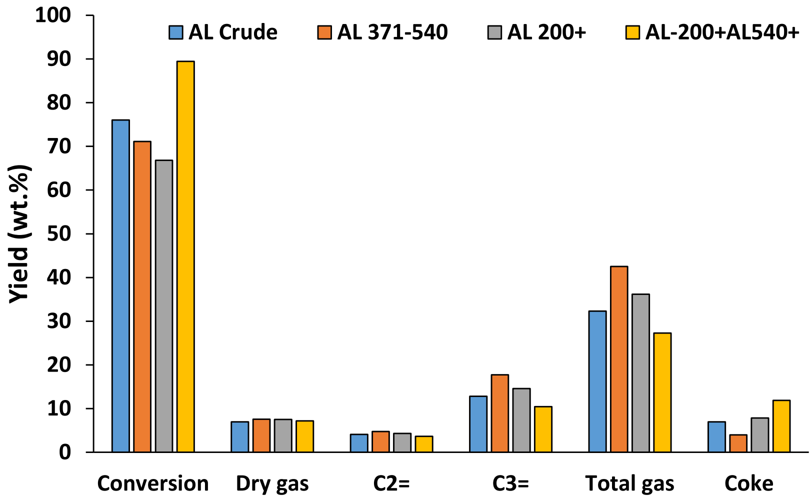


Figure 2. Effect of feed composition on conversion and product yields in the catalytic cracking reaction at 600 (°C), *C*/*O* = 5, and feed = 1.5 g.

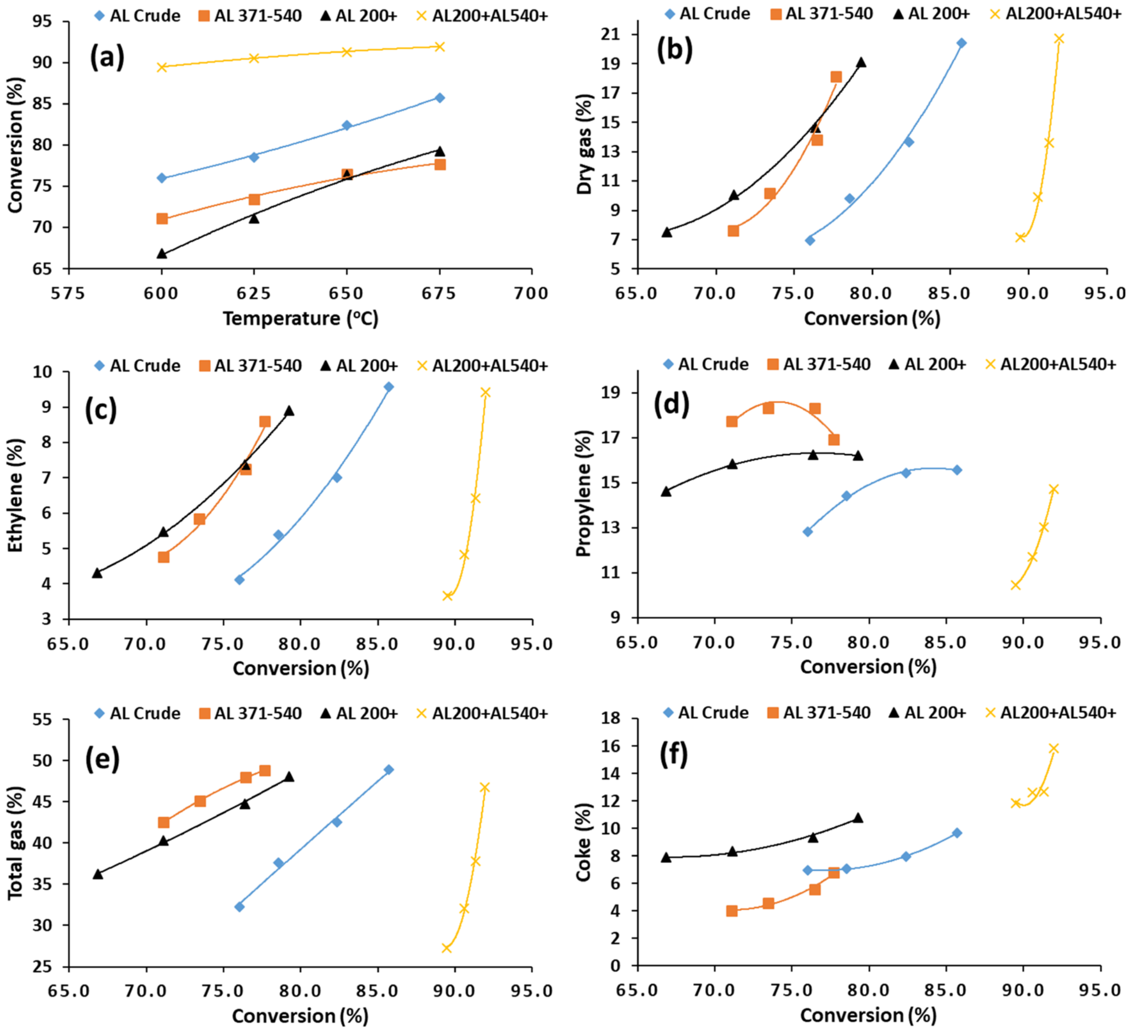


Figure 3. (a) Effect of temperature on conversion. (b−f) Evolution of product yields with conversion in the catalytic cracking of different crude oil samples at *C*/*O* = 5 and feed = 1.5 g.

the training data while maintaining a certain margin around the predicted values. The goal is to minimize the error between the predicted and actual target values while controlling the complexity of the model.

In the context of ensemble learning, SVR-E involves creating an ensemble of multiple standalone models (RT, LSQ-BOOST, GPR, and RLR). Each individual model is trained on the same data. The predictions from these individual models are then combined to make the final prediction.

Table 1. Descriptive Statistics of the Parameters Used for the Modeling

cal. set	temp. °C	feed	conversion (%)	total gas	dry gas	coke	ethylene
mean	637.50	256.38	79.89	41.18	12.71	8.88	6.43
min	600.00	0.00	66.83	27.34	6.96	4.02	3.66
max	675.00	455.50	91.94	48.86	20.75	15.82	9.57
SD	27.95	179.98	7.64	6.51	4.66	3.13	1.88
SN	16.00	16.00	16.00	16.00	16.00	16.00	16.00
val. set	temp. °C	feed	conversion (%)	total gas	dry gas	coke	propylene
mean	637.50	256.38	79.89	41.18	12.71	8.88	15.15
min	600.00	0.00	66.83	27.34	6.96	4.02	10.44
max	675.00	455.50	91.94	48.86	20.75	15.82	18.32
SD	27.95	179.98	7.64	6.51	4.66	3.13	2.21
SN	16.00	16.00	16.00	16.00	16.00	16.00	16.00

SD = standard deviation, min = minimum, max = maximum, SN = sample number, cal.= calibration, val. = validation.

2.8.3. Neural Network Ensemble (NNE). Nonlinear averaging is carried out using the neural ensemble approach (NNE) by training another neural network. The outputs of the single models are supplied into the input layer of the neural ensemble model, where each output is paired with a single input layer neuron. The tangent sigmoid is considered the activation function for the hidden and output layers in the neural ensemble approach, exactly as it is in the simple ANN. The network is trained using the back-propagation algorithm, and the ensemble network's optimal structure and epoch number are determined by trial and error.

2.8.4. Adaptive Neuro-Fuzzy Inference System Ensemble (ANFIS-E). An example of a hybrid model is one called ANFIS, which combines fuzzy logic with the strengths of neural networks. When complicated systems are modeled with uncertainty and imprecision, it is especially helpful. The rules of the ANFIS models are composed of a group of linguistic concepts and their corresponding membership functions. These guidelines are used to construct intermediate fuzzy outputs after fuzzifying the raw data. The relationships between the intermediate fuzzy outputs and the desired output are subsequently learned by using neural networks.

ANFIS-E involves creating an ensemble of multiple models to improve the predictive accuracy and robustness.

3. RESULTS AND DISCUSSION

3.1. Effect of Feed Composition on Conversion and Product Yields. The effect of feed composition on conversion and product yields was studied using four crude oil samples, Arabian Light (AL crude), AL 371−540, AL 200+, and AL-200+AL540+ blend, at a reactor temperature of 600 °C. As presented in Figure 2, the catalytic cracking of AL crude oil gives a conversion (%) of 76, whereas AL 371−540 and AL 200+ AL crude oil cuts give conversions (%) of 71 and 67, respectively. The AL crude oil has a wide range of low- to high-molecular-weight molecules, and since the enthalpy of the cracking reaction is relatively low for high-molecular-weight hydrocarbons, it makes sense that the conversion of AL crude oil is higher than that of AL 371−540, which is also higher than that of AL 200+. However, the blending of AL540+ (slurry) with AL 200+ to form AL-200+AL540+ increases the conversion to 89.5%.

This further confirms that the heavy hydrocarbon feedstock is easier to crack than the lighter ones, such as paraffins. The yield of dry gas in all of the cracked feeds is in the range of 7−7.6%, while the total gas yield is highest (42.50%) in AL 371−540. This implies that AL 371−540 has the highest yield of light olefins (ethylene = 4.8% and propylene = 17.7%). AL-200+AL540+ gives the lowest yield of total gas (27.3%) and the highest amount of coke. This is due to the large percentage of slurry in the feed.

3.2. Effect of Temperature on Conversion and Product Yields. The effect of temperature on conversion and product yields for the catalytic cracking of the four crude oil samples was studied by varying the reactor temperature from 600 to 675 °C. As presented in Figure 3a, a uniform increase was observed in conversion for AL crude and AL 200+. However, AL 371−540 showed a gradual increase from 600 to 650 °C and a slow increase at 675 °C. The AL-200+AL540+ feed showed a slow increase from 600 to 675 °C, and this is related to the initial high conversion (89.5%) recorded at 600 °C. However, the correlation of product yields with conversion gives a better understanding of the product yield progression with temperature. Thus, the plot of dry gas yield versus conversion (Figure 3b) shows a relatively linear relationship for AL crude oil and AL 200+; however, AL 371−540 and especially AL-200+AL540+ show an exponential relationship. This implies that conversions of less than 73 and 90% are required to keep the dry gas level below 10% in AL 371−540 and AL-200+AL540+ feeds, respectively. However, this will affect the yield of ethylene, as shown in Figure 3c, which is nearly 9% in both AL 371−540 and AL-200+AL540+ cracking at their highest conversions of 78 and 92%, respectively. Interestingly, the yield of propylene in the cracking of the AL 371−540 feed gives a plateau at a conversion below 73%, with the highest propylene yield of 18.3%, while AL-200+AL540+ cracking shows a continuous increase in propylene yield, although the yield is only 14.7% at 92% conversion. It follows that at higher conversion (>73%) during cracking of the AL 371−540 feed, some of the propylene undergoes further reaction via the hydrogen transfer mechanism to oligomerize and aromatize. Similarly, both AL crude and AL 200+ feeds show maxima in the plot of propylene versus conversion (Figure 3d) at conversions of 82 and 76%, respectively. Therefore, to maintain the maximum yield of propylene in all of the feeds, a certain conversion threshold must not be exceeded, and this can be correlated to the reactor temperature as well. The total gas yield shows pretty much the same trend for all of the feed compositions, increasing trend with conversion; however, the AL-200+AL540+ feed shows an exponential increase in the total gas yield with conversion.

3.3. Modeling the Yield of Light Olefins in the Cracking of Crude Oil Samples. The observed trends in conversion and product yields with different crude oil cuts and at different reactor temperatures indicated that the maximum yields of ethylene and propylene at the maximum conversion are

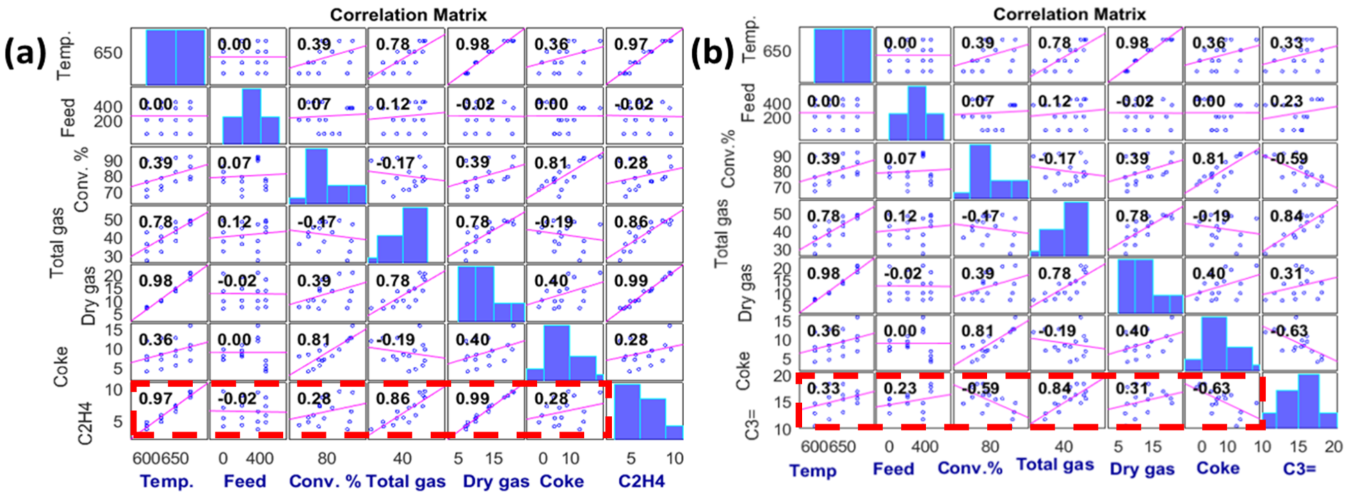


Figure 4. Input−output feature extraction for (a) ethylene and (b) propylene.

achievable through modeling of the crude oil cracking. A successful model will enable effective prediction of product yields under different variable conditions, and in this study, the products of interest are ethylene and propylene. Table 1 presents the descriptive statistics of the parameters used in the modeling, indicating their mean values, standard deviation, sample variance, minimum, and maximum. For instance, ethylene showed a maximum concentration value of 9.57 and a minimum concentration value of 3.66. This indicates that the concentration range is 5.91. More so, the descriptive statistics equally showed that propylene had a mean value of 6.43 and a standard deviation of 1.94.

3.3.1. Input−Output Feature Extraction. Among the basic fundamentals of any chemometric-based data intelligence technique is the feature selection of input variables. This is equally applicable in the area of catalytic cracking, which is associated with various chemical and physical phenomena. However, to date, there has been no single technique that has been proven to be the best in selecting appropriate input candidates in chemometrics feature extraction. Generally, most of the techniques, especially those related to a descriptor-based method for feature selection using the clustering calculation method, are considered to be the fundamental and primary steps but equally depict some limitations in choosing the suitable composition. Clustering is a technique in unsupervised machine learning that involves grouping similar data points together into clusters based on certain features or characteristics. One of the most commonly used clustering algorithms is *K*-means. Some of the common limitations of clustering calculations as a feature selection approach consist of sensitivity to initialization, hierarchical clustering complexity, assumption of equal cluster sizes and shapes, etc. Therefore, the current work employs the use of a linear matrix input−output-based feature extraction technique (see Figure 4).

The combinations of input variables were selected according to the correlational analysis for the dependent−independent variable relation as the feature selection. Therefore, the variables were classified into three different classes for each of the dependent variables (ethylene and propylene). For ethylene as the target, the combinations were developed as M1 (feed, conversion %, coke), M2 (temperature, dry gas, total gas), and M3 (temperature, feed, conversion %, total gas, dry gas, coke). Moreover, for propylene, the combinations were developed as M1 (temperature, feed, dry gas), M2 (conversion %, total gas, coke), and M3 (temperature, feed, conversion %, total gas, dry gas, coke). It is significant to mention that the idea of feature selection is well established in various fields of chemometrics.[36−38] As mentioned earlier, the data used in this experiment were obtained from the ACE unit experiments conducted in the laboratory. To ensure the accuracy and reliability of the results, the data were divided into two subsets called the train−test split after loading the data in MATLAB (2020a). In a train−test split, the original data set is divided into two parts: a training set and a testing (or validation) set. The training set is used to train the machine learning model, while the testing set is used to evaluate the model's performance on unseen data. Typically, a larger portion of the data are allocated to the training set (e.g., 70−80%) and a smaller portion to the testing set (e.g., 20−30%). The split is usually done randomly to ensure that the data in both sets are representative of the overall data set. Based on the current study, the splitting was done as a training data set consisting of 75% of the data and a validation set consisting of the remaining 25%. These subsets were then utilized in the development of models using both single and EL models in order to improve the accuracy of the results. By utilizing a diverse range of modeling approaches, we were able to gain a more comprehensive understanding of the data and increase the overall robustness of our findings.

3.3.2. Results of the Single Models. The current section presents and describes the obtained results from the standalone ML-based techniques and the EL computational technique for the prediction of ethylene and propylene concentration yields obtained from the catalytic cracking of crude oil fractions. The prediction of the ML techniques was made on MATLAB (2020a) MathWorks, Inc., United States. The performance of the ML paradigms is presented in Tables 2 and 3 for ethylene and propylene yields, respectively. Moreover, three different grading metrics, namely, coefficient of determination (R^2), Pearson correlation coefficient (PCC), and mean square error (MSE), were used for evaluating the performance of the models. R^2 is a dimensionless value between 0 and 1, which represents the proportion of the variance in the simulated values (Y) that are simulated by the independent variable(s)/experimental values (X). It does not have a unit because it is a relative measure of goodness of fit, indicating the percentage of variation explained. Similarly, the PCC is also a dimensionless value

Table 2. Performance of the Single Models for the Prediction of Ethylene Concentration

	R^2	calibration PCC	MSE (wt %)
RT-M1	0.42	0.65	2.05
RT-M2	0.75	0.87	0.87
RT-M3	0.99	0.99	0.04
LSQ-BOOST-M1	0.07	0.27	3.29
LSQ-BOOST-M2	0.80	0.90	0.70
LSQ-BOOST-M3	0.81	0.90	0.66
GPR-M1	0.97	0.99	0.09
GPR-M2	0.99	0.99	0.05
GPR-M3	**1.00**	**1.00**	**0.00**
RLR-M1	0.09	0.29	3.24
RLR-M2	0.97	0.99	0.10
RLR-M3	0.99	0.99	0.04
		validation	
RT-M1	0.38	0.62	2.06
RT-M2	0.71	0.84	0.88
RT-M3	0.93	0.96	0.05
LSQ-BOOST-M1	0.05	0.23	3.30
LSQ-BOOST-M2	0.80	0.89	0.71
LSQ-BOOST-M3	0.80	0.90	0.67
GPR-M1	0.96	0.98	0.10
GPR-M2	0.98	0.99	0.06
GPR-M3	**0.99**	**0.99**	**0.01**
RLR-M1	0.08	0.27	3.25
RLR-M2	0.96	0.98	0.11
RLR-M3	0.96	0.98	0.05

Note: bold represents the best performing model.

Table 3. Performance of the Single Models for the Prediction of Propylene Concentration

		calibration	
	R^2	PCC	MSE (wt %)
RT-M1	0.73	0.86	1.30
RT-M2	0.80	0.89	0.98
RT-M3	0.81	0.90	0.95
LSQ-BOOST-M1	0.01	0.07	4.86
LSQ-BOOST-M2	0.56	0.75	2.16
LSQ-BOOST-M3	0.87	0.93	0.62
GPR-M1	0.01	0.06	4.89
GPR-M2	0.99	1.00	0.03
GPR-M3	**1.00**	**1.00**	**0.00**
RLR-M1	0.16	0.40	4.10
RLR-M2	0.95	0.97	0.26
RLR-M3	0.98	0.99	0.10
		validation	
RT-M1	0.71	0.84	1.33
RT-M2	0.71	0.84	1.01
RT-M3	0.80	0.89	0.98
LSQ-BOOST-M1	0.01	0.06	4.90
LSQ-BOOST-M2	0.55	0.74	2.19
LSQ-BOOST-M3	0.85	0.92	0.65
GPR-M1	0.01	0.06	4.92
GPR-M2	0.99	1.00	0.07
GPR-M3	**1.00**	**1.00**	**0.04**
RLR-M1	0.19	0.44	4.13
RLR-M2	0.94	0.97	0.29
RLR-M3	0.98	0.99	0.13

Note: bold represents the best performing model.

that measures the strength and direction of a linear relationship between two continuous variables. An R^2 or PCC value that is very close to 1 indicates that the model fits the experimental data accurately. The unit of MSE depends on the unit of the data being analyzed, as it is wt % for this study. It represents the average squared difference between the observed values and the predicted values, so it maintains the units of the dependent variable. An MSE value that is very close to 0 indicates that the model fits the experimental data accurately.

Table 2 depicts the quantitative performance of the single models. Based on the R^2, PCC, and MSE values, it was observed that M2 and M3 input combinations outperformed M1 in all four standalone techniques. Even though GPR outperformed all of the models in both the training and validation phases, RLR equally demonstrated outstanding performance. This can be attributed to the robust ability of the linear technique over other linear classical techniques such as interactive linear regression (ILR), stepwise linear regression (SWLR), and multivariate regression (MVR).[39] Another reason why RLR can give higher performance in prediction is that it is less affected by outliers than traditional linear regression. Outliers can have a significant impact on the results of linear regression, as they can cause the estimated regression line to be skewed or biased toward the outliers.[40] RLR uses robust estimation methods, such as M-estimation or Huber's method, which are less affected by outliers and can provide more accurate estimates of the regression coefficients. Also, RLR can give higher performance in prediction owing to the fact that it is more robust to violations of normality assumptions.[33] Traditional linear regression assumes that the errors in the model are normally distributed, but in practice, this assumption may not hold. RLR uses estimation methods that are less sensitive to non-normality in the data, which can lead to more accurate predictions.

Similarly, the results indicate that the models failed to accurately model the ethylene concentration in certain instances. Therefore, there is a need to employ advanced computational techniques, such as ensemble paradigms, hybrid methods, and metaheuristic approaches, to enhance the performance of the individual models. Furthermore, the performance of these individual models can be visualized using both scatter plots and response plots, as shown in Figures 5 and 6. Figure 5 represents the response plot, a data visualization technique used to display the relationship between two numerical variables. Each data point on the plot corresponds to an individual observation in the data set, with its position determined by the values of the two variables being plotted. Figure 6 presents a scatter plot in the form of a time series, which is another data visualization method used to display data points arranged chronologically over time. Time series plots are particularly valuable for visualizing how a variable changes over time and identifying trends, patterns, seasonality, and potential anomalies. Based on the observed graphical trends, it is evident that there is a weak correlation between the simulated and experimental ethylene yields. Hence, most of the model combinations depict lower performances with the exception of M3 and M2 combinations for some instances, just as in RT-M3 with R^2 values of 0.99 and 0.93, LSQ-BOOST-M2 (0.80, 0.80), LSQ-BOOST-M3 (0.81, 0.80), GPR-M1 (0.97, 0.96), GPR-M2 (0.99,0.98), GPR-M3 (1.00,0.99), RLR-M2 (0.97, 0.96), and RLR-M3 (0.99, 0.96) in both the training and testing phases respectively. Moreover, based on the undulation depicted by the time series plot that changes over time, it indicates that the M3 model combination showed higher performance than M1 and

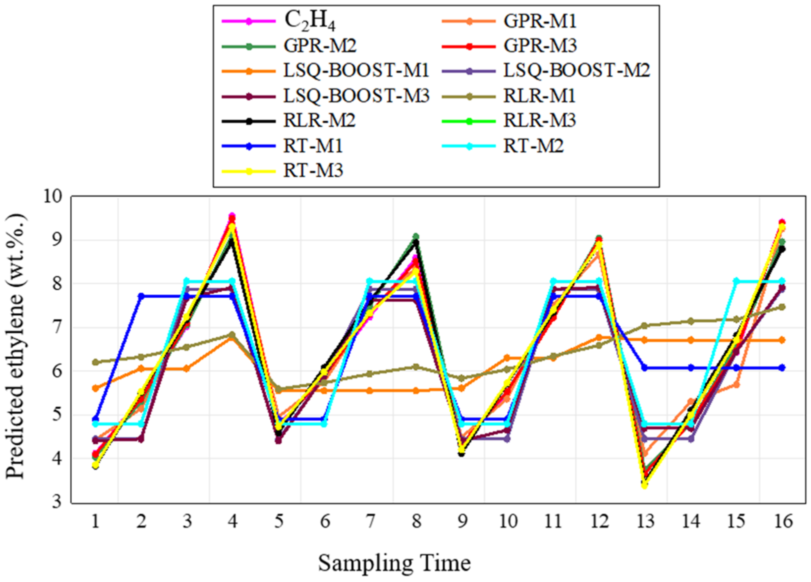


Figure 5. Time series plot performance of the single models for ethylene concentration modeling.

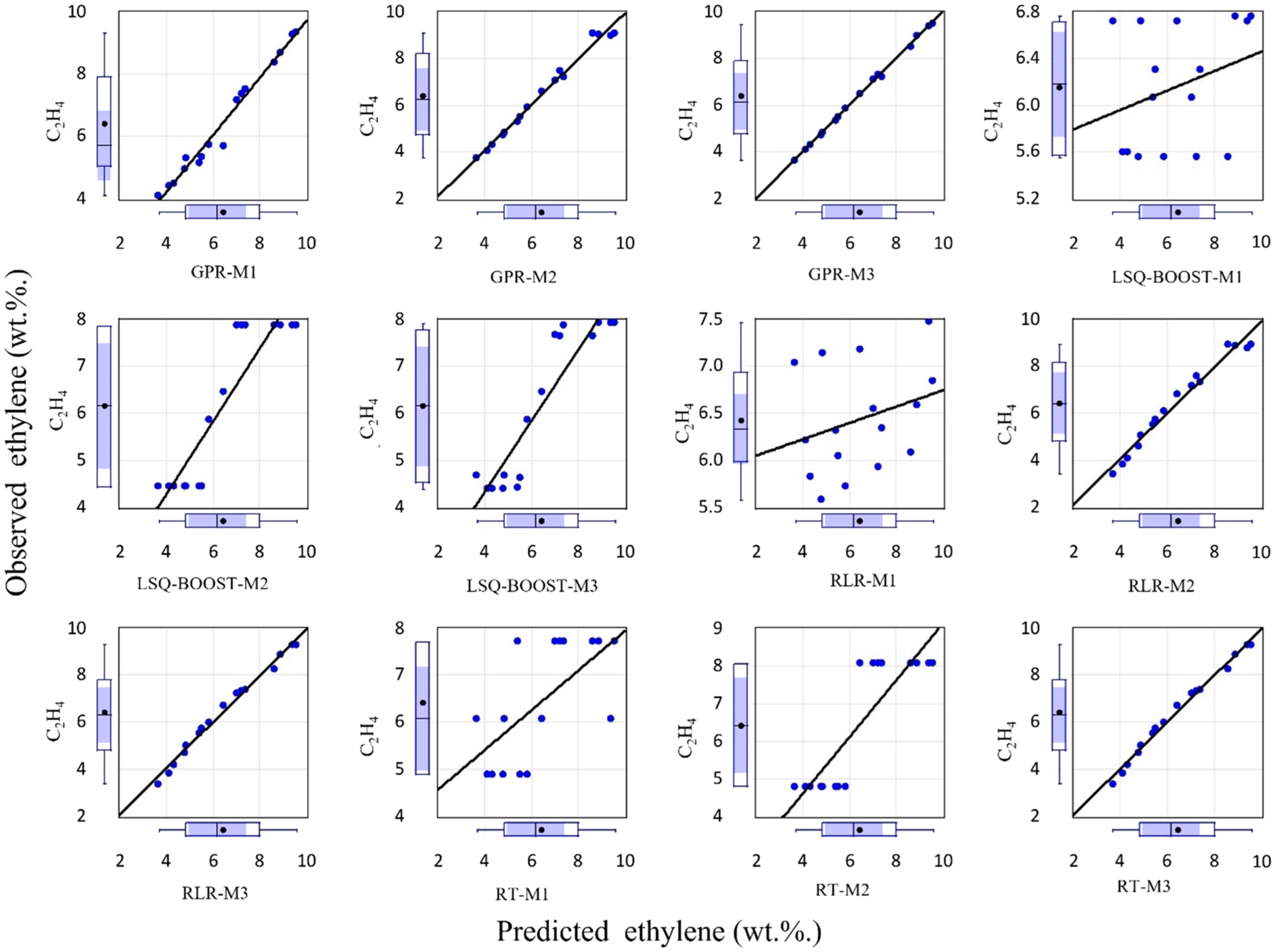


Figure 6. Scatter plot performance of the single models for ethylene concentration modeling.

M2. Moreover, among the four models, GPR showed that the target was captured more than other models.

Furthermore, Table 3 indicates the numerical performance of the standalone paradigms. The performance results indicate the inability of some models based on different input combinations to model the prediction performance of the ethylene concentration. According to Abba et al.,[41] for any chemometric-based approach, a data-driven model to be accepted should depict at least 80% performance fitness in both the training and testing stages.

Therefore, this study employed the application of robust ensemble ML techniques in order to improve the prediction efficiency of the single models. Moreover, the prediction performance can equally be demonstrated graphically using the time series and scatter plots, as indicated in Figures 7 and 8, respectively. Figure 7 is a graphical illustration that demonstrates the comparative performance of the standalone models. Moreover, the graph indicates that GPR-M3 outperformed all of the other model combinations. If GPR-M3 model-based predictions closely overlap with observed values on a graph, it suggests that its predictions are highly accurate, capturing the underlying data patterns effectively. This agreement indicates reduced error margins (MSE = 0.000 wt %), implying superior performance to other models. Consistency in this overlap across all data points demonstrates the model's reliability, while a similar performance on unseen data would signify its ability to generalize well. However, it is obvious that perfect agreement, especially on training data, could hint at overfitting, where the model might be too fitted to be training data nuances, risking its performance on new data. In essence, while GPR-M3′s overlap with observed data emphasizes its potential strength, ensuring that it performs well on unfamiliar data is vital to confirm its broad applicability and to prevent overfitting concerns, as recommended.

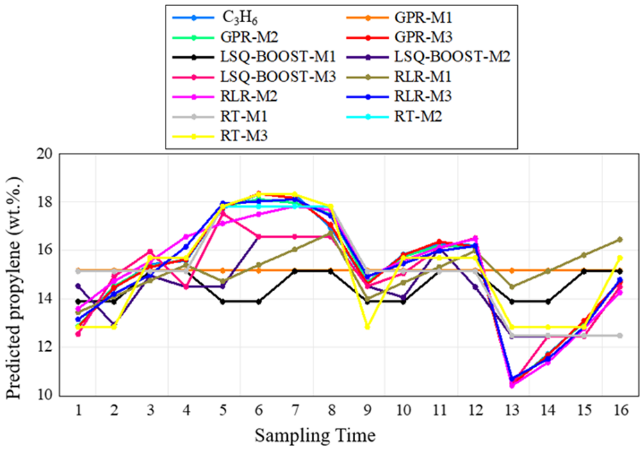


Figure 7. Time series plot performance of the single models for propylene concentration modeling.

3.3.3. Results of the Ensemble ML Paradigms. Owing to the accuracy being lower in some instances depicted by the standalone models in modeling ethylene and propylene concentrations, the novel ensemble technique was proposed. Hence, the quantitative performance of the ensemble ML techniques (GRNN-E, SVR-E, NNE, and ANFIS-E) for modeling ethylene is equally shown in Table 4. The PCC is a

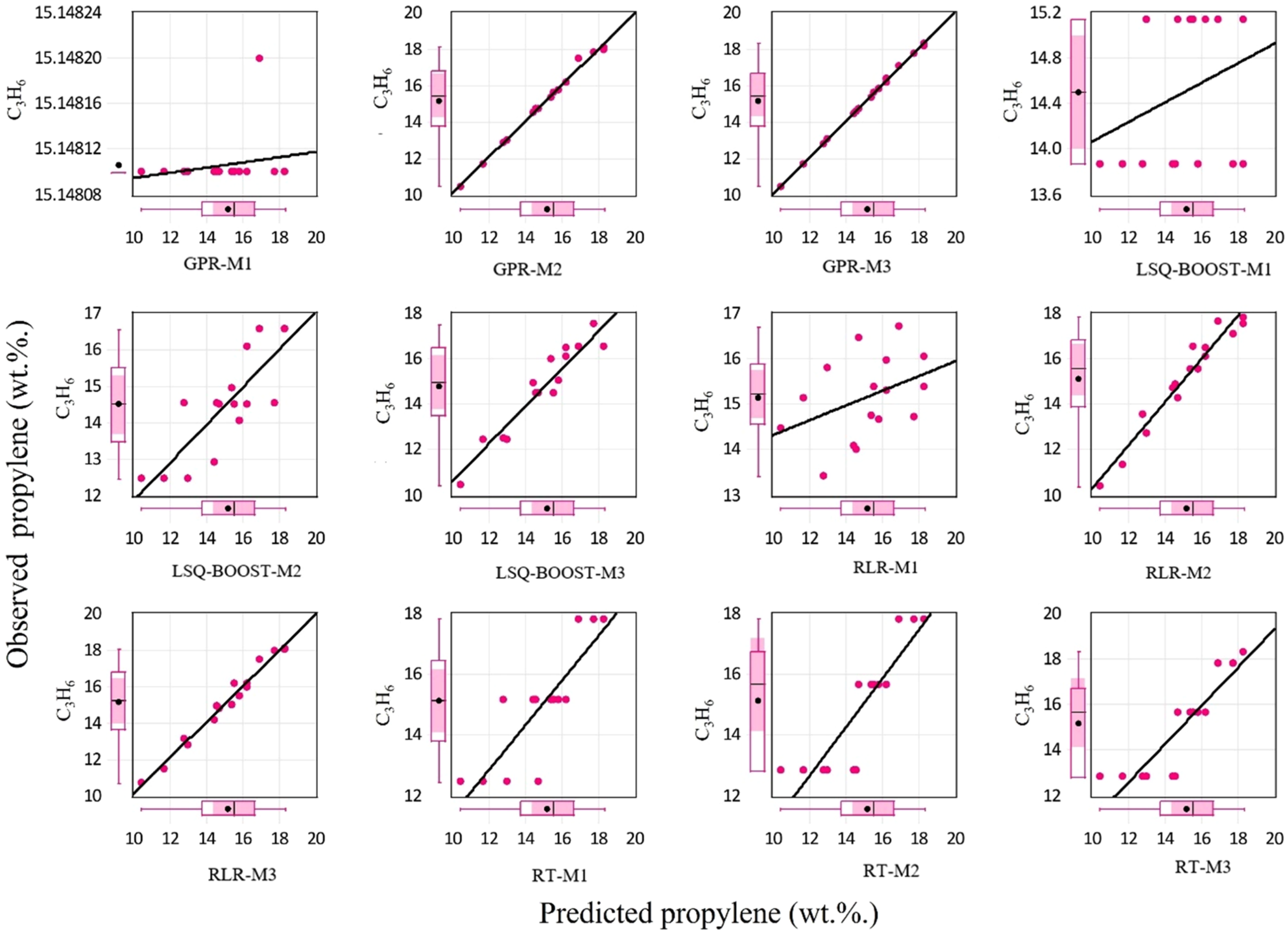


Figure 8. Scatter plot performance of the single models for propylene concentration modeling.

Table 4. Results of the Ensemble Paradigms for the Prediction of Ethylene Concentration

	calibration		
	R^2	PCC	MSE (wt %)
GRNN-E	0.99	1.00	0.00
SVR-E	0.99	1.00	0.01
NNE	0.99	1.00	0.00
ANFIS-E	**1.00**	**1.00**	**0.00**
	validation		
GRNN-E	0.99	0.99	0.01
SVR-E	0.99	0.99	0.01
NNE	0.99	0.99	0.01
ANFIS-E	**1.00**	**1.00**	**0.01**

Note: bold represents the best performing model.

statistic that measures the strength and direction of the linear relationship between two continuous variables. It quantifies how well variation in one variable can be predicted by variation in another variable. In the current study, PCC was used to show the relationship between the simulated and experimental studies. Also, as indicated in Table 4, all of the four ensemble paradigms depict higher PCC values in both the calibration and validation phases. For instance, the best performing technique ANFIS-E presents a PCC value of 1.00 in both the training and validation phases, whereas GRNN-E, SVR-E, and NNE show PCC values of 1.00 in the training and 0.99 in the validation stage in ethylene concentration modeling.

The performance accuracy of the ensemble paradigms presented in Table 3 demonstrates the robust application of the novel ensemble technique over the traditional single approaches. Hence, the techniques were able to dramatically enhance the performance of the models. To compare the performance of the current research with recent work done on the technical published literature, Kawai et al.[42] reported an AI hybrid-based model to improve the catalyst makeup rate and increase the product yield during real-time operation by creating a reaction model. This involves developing a method for evaluating catalyst activity as well as integrating the C/O ratio to assess reaction performance. To further enhance the process, a yield prediction model is incorporated into state-of-the-art digital technologies. Hence, the AI hybrid-based model depicts MSE values ranging from 0.06 to 0.21 wt %, while the developed EL in our research depicts MSE values that range from 0.0000 to 0.0057 wt % in the training and from 0.0065 to 0.0122 wt % in the validation step. This indicates that the performance of our EL models outperformed theirs. Moreover, Yang et al.[43] reported the implementation of different deep learning techniques for the prediction of fluid catalytic cracking (FCC). The comparative performance of these models indicates that the models depict MSE values ranging from 0.0062 to 0.0144 wt % in the training phase and from 0.0066 to 0.0146 wt % in the testing phase. Hence, the performance of the developed EL models has outperformed the performance of deep learning techniques.

Moreover, the performance of the ensemble paradigms can be demonstrated graphically based on their error performance (see Figure 9). Moreover, the fitness comparative performance of the ensemble paradigms can equally be depicted graphically based on the scatter and time series visualization plots (Figures 10 and 11, respectively).

The MSE of the calibration set in the experiment significantly exceeded that of the validation set, pointing to a likely case of model overfitting. This overfitting implies that the model, while appearing to perform exceedingly well on the training data, may actually just be memorizing the data rather than generalizing from it, which would result in poor performance on new, unseen data. This situation could potentially be attributed to the methodology employed during the sampling phase. This approach can often lead to a lack of diversity and variation within the sample set, which, in turn, can contribute to overfitting. When a model is trained on such a data set, it might become overly tailored to that specific set of data, failing to generalize and perform well on new data. This is because the model was overly optimized for the specific characteristics of the training data. It is crucial, therefore, to reassess the sampling approach, possibly introducing replication and ensuring a more diverse and representative sample set to improve the robustness and generalizability of the model, thereby mitigating the issue of overfitting.

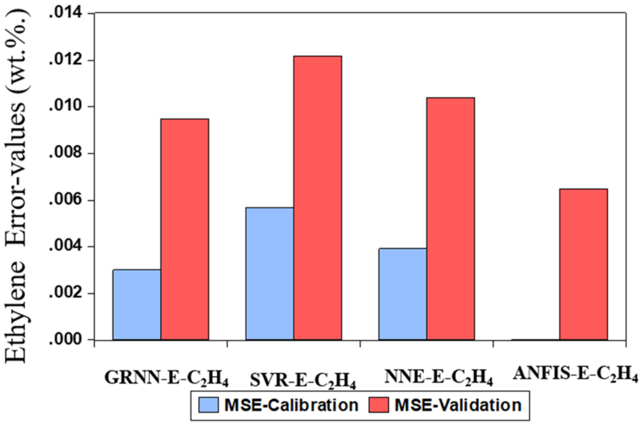


Figure 9. Error performance of the ensemble paradigms for ethylene concentration modeling.

Figure 11 demonstrates the comparative performance of ensemble techniques toward modeling the ethylene concentration. The graphical illustration indicates the robust ability of all four ensemble paradigms with satisfactory and reliable accuracy, as demonstrated in Table 5.

Furthermore, Table 5 indicates the numerical performance of the ensemble paradigms. The performance results indicate the robust abilities of ensemble ML in modeling the prediction performance of the propylene yield. Therefore, this study employed the application of a robust ensemble ML technique in order to improve the prediction efficiency of the single models. As shown in Table 5, there is too much difference in the MSE for calibration and validation, which can be attributed to the different complexities of the ensemble machine learning approach, which utilizes the results from the single ML techniques as the input variables. This can lead to a phenomenon called “data mismatch,” that is, if the training data and testing data come from different distributions or have significant differences in their characteristics, the model may struggle to generalize. It might perform well on the training data but poorly on the testing data, leading to a high MSE in testing.

Moreover, the prediction performance of the models by the four modeling algorithms can equally be demonstrated graphically using the time series and scatter plots, as indicated in Figures 12 and 13, respectively.

Besides, the performance accuracy of the models can equally be checked graphically using the bar chart based on the respective MSE performance of the models, as indicated in

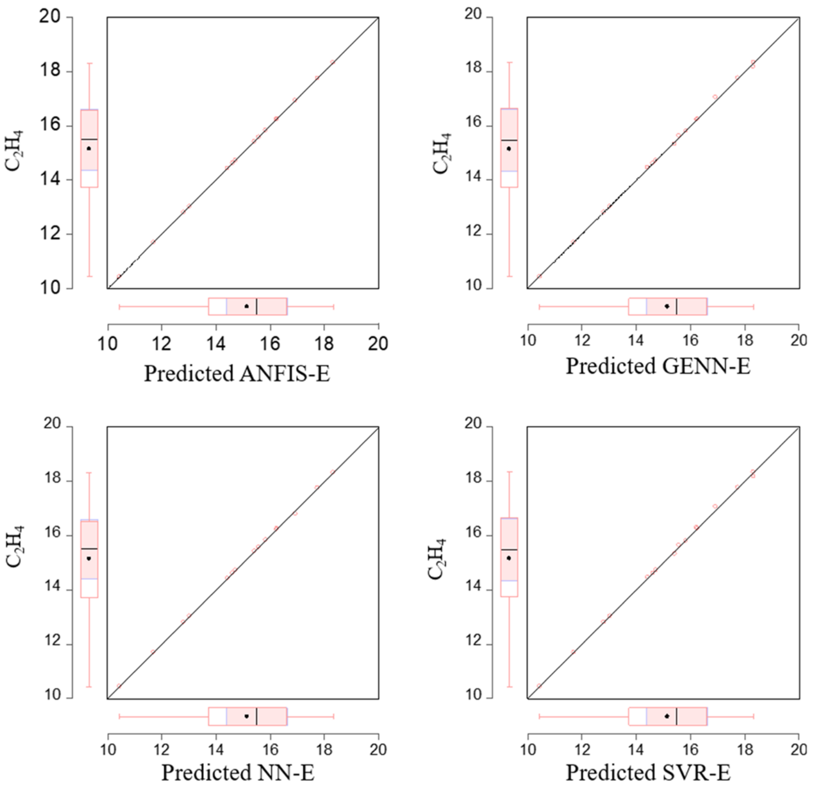


Figure 10. Scatter plot for ethylene concentration modeling.

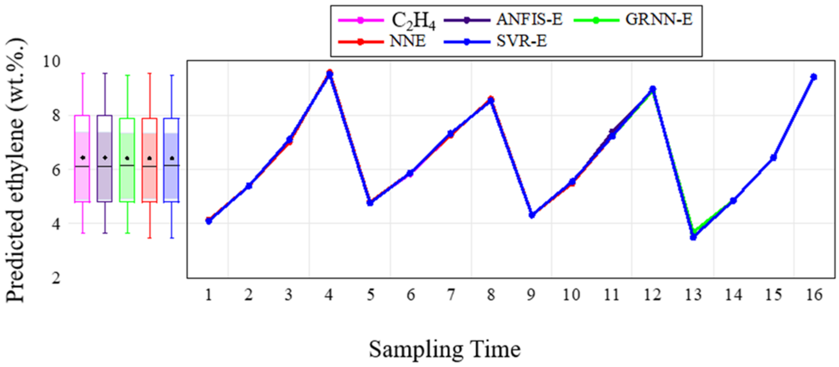


Figure 11. Time series plot of the ensemble paradigms for ethylene concentration modeling.

Table 5. Performance of the Ensemble Paradigms for the Prediction of Propylene Concentration

	calibration		
	R^2	PCC	MSE (wt %)
GRNN-E	0.99	1.00	0.01
SVR-E	0.99	1.00	0.01
NNE	0.99	1.00	0.00
ANFIS-E	**1.00**	**1.00**	**0.00**
	validation		
GRNN-E	0.99	1.00	0.09
SVR-E	0.99	1.00	0.09
NNE	0.99	1.00	0.08
ANFIS-E	**1.00**	**1.00**	**0.08**

Note: bold represents the best performing model.

Figure 14. According to Figure 14, it can be understood that the MSE calibration performance depicts relatively higher performance than the MSE validation, which is attributed to the data mismatch, as illustrated previously. Furthermore, the overall comparative performance depicted in Figure 14 demonstrated that ANFIS-E and NNE techniques showed comparatively similar performances and outperformed SVR-E and GRNN-E ensemble machine learning techniques. The MSE values generated through the calibration stage were much higher than the MSE values for the validation set, which signifies the overfitting of the model owing to the wide and significant range between the MSE values of the two phases. This can be attributed to the fact that sampling was done without replication and less sample variation.

Zhu et al.[44] employed artificial neural networks (ANNs) as a machine learning (ML) technique for modeling and establishing protocols to understand the impact of catalysts and temperature on propene and ethylene production through *n*-pentane cracking. The performance of the ML model was assessed using two different metrics: R^2 and MSE. Consequently, the results of their study revealed lower R^2 values and higher MSE

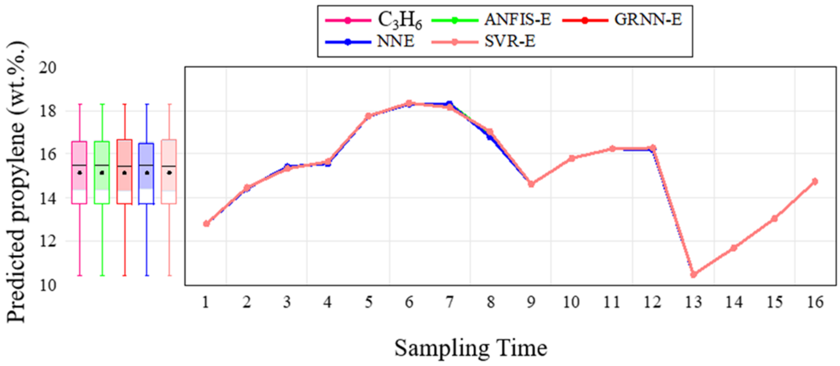


Figure 12. Time series plot of the ensemble paradigms for propylene concentration modeling.

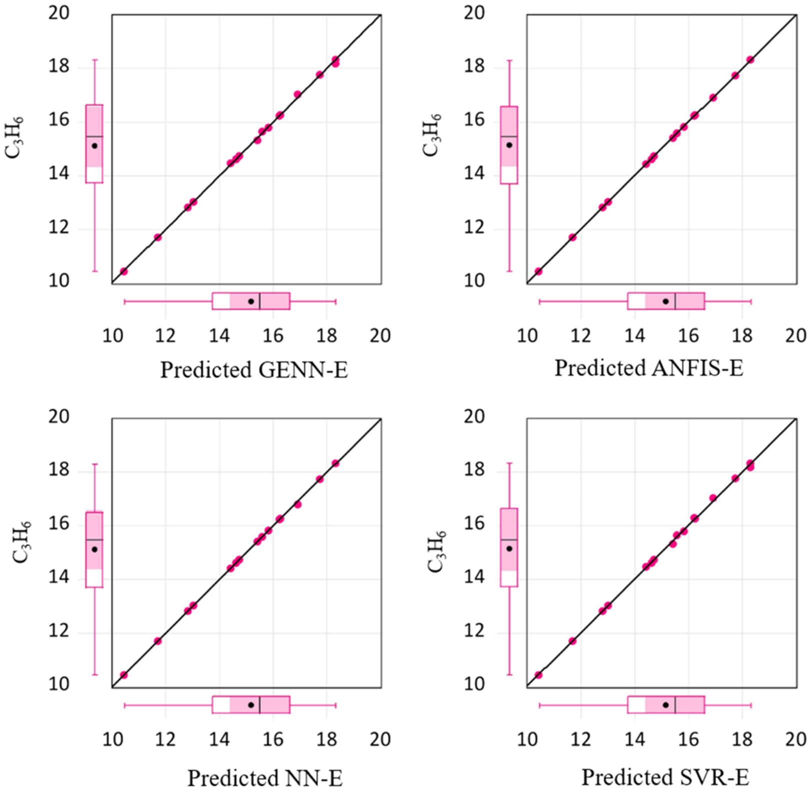


Figure 13. Scatter plot for propylene concentration modeling.

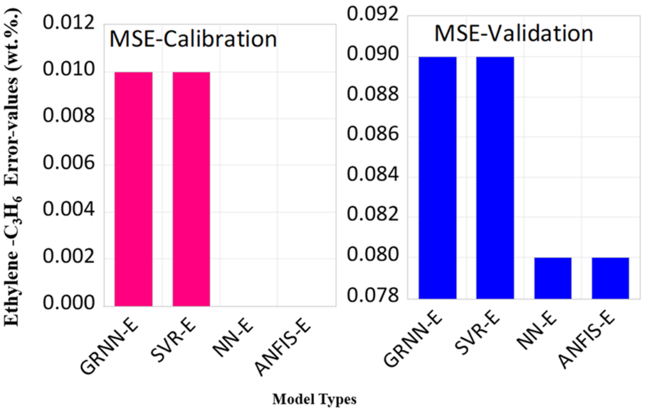


Figure 14. Error performance of the ensemble paradigms for propylene concentration modeling.

values compared to the ones generated by the EL algorithms developed in our study.

Also, Zhu and Wang[45] integrated the use of the cuckoo search algorithm and Elman neural networks (ENNs) for modeling the reactor−regenerator system as an important factor in the fluid catalytic cracking unit using variables. The obtained results present the performance values of PCC metrics ranging from 0.9536 to 0.9980 in both the training and testing phases for the best model combination, whereas the performance of the EL algorithms for the prediction of propylene concentration reported in the current study depicted PCC values ranging from 0.9996 to 1.000. This clearly indicates that the EL algorithms developed in our study depict robust and promising performance compared to other techniques depicted in the literature.

4. CONCLUSIONS

The catalytic cracking of AL crude oil and three crude oil fractions has been carried out in an advanced catalyst evaluation (ACE) reactor unit, and the effect of feed composition and reactor temperature on conversion and product yields was studied. The data sets generated were employed in AI-based models' development for the prediction of yield of light olefin products, particularly ethylene and propylene. Four single-model AI techniques and four ensemble paradigms were used to develop the prediction models, and five different variables, namely, temperature, feed type, feed conversion, total gas, and coke, were imputed. Feature selection was conducted using correlation matrix analysis, and performance metrics were used to evaluate the prediction performance of the developed models in both calibration and validation stages. The performance metrics are the coefficient of determination (R^2), Pearson correlation coefficient (PCC), and mean square error (MSE). The single models depict very low R^2 and PCC values (as low as 0.07) and very high MSE values (up to 4.92 wt %) for M1 and M2 in both calibration and validation phases, while the ensemble ML models show R^2 and PCC values of 0.99−1 and an MSE value of 0.01 wt % for M1, M2, and M3 input combinations in both calibration and validation phases. Therefore, ethylene and propylene yield prediction performance were more accurate in the case of ensemble paradigm-developed models, with an estimated performance improvement of 58 and 62% in the calibration and validation phases, respectively, in comparison with the standalone models.

■ AUTHOR INFORMATION

Corresponding Author

Abdulkadir Tanimu − *Center for Refining and Advanced Chemicals, Research Institute, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia;* orcid.org/0000-0002-7541-0042; Email: abdulkadir.tanimu.1@kfupm.edu.sa

Authors

A. G. Usman − *Department of Analytical Chemistry, Faculty of Pharmacy, Near East University, 99138 Nicosia, Turkey; Operational Research Centre in Healthcare, Near East University, 99138 Nicosia, Turkish Republic of Northern Cyprus*

S. I. Abba − *Interdisciplinary Research Center for Membrane and Water Security, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*

Selin Isik − *Department of Analytical Chemistry, Faculty of Pharmacy, Near East University, 99138 Nicosia, Turkey*

Abdullah Aitani − *Center for Refining and Advanced Chemicals, Research Institute, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia;* orcid.org/0000-0001-5071-4034

Hassan Alasiri − *Center for Refining and Advanced Chemicals, Research Institute, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; Department of Chemical Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* orcid.org/0000-0003-4043-5677

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.3c05227

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The financial support of the Deanship of Research Oversight and Coordination, King Fahd University of Petroleum and Minerals under Project INRC2214 is highly appreciated.

■ REFERENCES

(1) Nagpal, S. Crude Oil Conversion to Chemicals-IHS Markit Process Economics Program. 2021.

(2) Zhou, X.; Sun, Z.; Yan, H.; Feng, X.; Zhao, H.; Liu, Y.; et al. Produce petrochemicals directly from crude oil catalytic cracking, a techno-economic analysis and life cycle society-environment assessment. *J. Cleaner Prod.* **2021**, *308*, No. 127283.

(3) Petrochemical Market Size to Worth Around US$ 729 Bn by 2030 n.d. https://www.precedenceresearch.com/petrochemical-market (accessed December 30, 2021).

(4) Akah, A.; Williams, J.; Ghrami, M. An Overview of Light Olefins Production via Steam Enhanced Catalytic Cracking. *Catal. Surv. Asia* **2019**, *23*, 265−276.

(5) Tanimu, A.; Tanimu, G.; Alasiri, H.; Aitani, A. Catalytic Cracking of Crude Oil: Mini Review of Catalyst Formulations for Enhanced Selectivity to Light Olefins. *Energy Fuels* **2022**, *36*, 5152−5166.

(6) Corma, A.; Corresa, E.; Mathieu, Y.; Sauvanaud, L.; Al-Bogami, S.; Al-Ghrami, M. S.; Bourane, A. Crude oil to chemicals: light olefins from crude oil. *Catal. Sci. Technol.* **2017**, *7*, 12−46.

(7) Qureshi, Z. S.; Siddiqui, M. A.; Tanimu, A.; Aitani, A.; Akah, A. C.; Xu, Q.; AlHerz, M. Steam catalytic cracking of crude oil over novel hierarchical zeolite−containing mesoporous silica−alumina core-shell catalysts. *J. Anal. Appl. Pyrolysis* **2022**, *166*, No. 105621.

(8) Jermy, B. R.; Tanimu, A.; Siddiqui, M. A.; Qureshi, Z. S.; Aitani, A.; Akah, A.; et al. Crude oil conversion to chemicals over green synthesized ZSM-5 zeolite. *Fuel Process. Technol.* **2023**, *241*, No. 107610.

(9) Maadhah, A. G.; Abul-Hamayel, M.; Aitani, A. M.; Ino, T.; Okuhara, T. Down-flowing FCC reactor. *Oil Gas J.* **2000**, *98*, 66−68.

(10) Alabdullah, M.; Rodriguez-Gomez, A.; Shoinkhorova, T.; Dikhtiarenko, A.; Chowdhury, A. D.; Hita, I.; et al. One-step conversion of crude oil to light olefins using a multi-zone reactor. *Nat. Catal.* **2021**, *4*, 233−241.

(11) Usman, A.; Siddiqui, M. A. B.; Hussain, A.; Aitani, A.; Al-Khattaf, S. Catalytic cracking of crude oil to light olefins and naphtha: Experimental and kinetic modeling. *Chem. Eng. Res. Des.* **2017**, *120*, 121−137.

(12) Al-Khattaf, S.; Saeed, M. R.; Aitani, A.; Klein, M. T. Catalytic Cracking of Light Crude Oil to Light Olefins and Naphtha over E-Cat and MFI: Microactivity Test versus Advanced Cracking Evaluation and the Effect of High Reaction Temperature. *Energy Fuels* **2018**, *32*, 6189−6199.

(13) Blay, V.; Louis, B.; Miravalles, R.; Yokoi, T.; Peccatiello, K. A.; Clough, M.; Yilmaz, B. Engineering zeolites for catalytic cracking to light olefins. *ACS Catal.* **2017**, *7*, 6542−6566.

(14) Kotrel, S.; Knözinger, H.; Gates, B. C. The Haag−Dessau mechanism of protolytic cracking of alkanes. *Microporous Mesoporous Mater.* **2000**, 35−36, 11−20.

(15) Gbadago, D. Q.; Moon, J.; Kim, M.; Hwang, S. A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and genetic algorithm: A case of butadiene synthesis. *Chem. Eng. J.* **2021**, *409*, No. 128163.

(16) Bollas, G. M.; Papadokonstantakis, S.; Michalopoulos, J.; Arampatzis, G.; Lappas, A. A.; Vasalos, I. A.; Lygeros, A. A Computer-Aided Tool for the Simulation and Optimization of the Combined HDS−FCC Processes. *Chem. Eng. Res. Des.* **2004**, *82*, 881−894.

(17) Bollas, G. M.; Papadokonstadakis, S.; Michalopoulos, J.; Arampatzis, G.; Lappas, A. A.; Vasalos, I. A.; Lygeros, A. Using hybrid

neural networks in scaling up an FCC model from a pilot plant to an industrial unit. *Chem. Eng. Process.: Process Intensif.* **2003**, *42*, 697−713.

(18) Kawai, E.; Sato, H.; Furuichi, K.; Takatsuka, T.; Yoshioka, T. Maximizing margins and optimizing operational conditions for residue fluid catalytic cracking with an artificial intelligence hybrid reaction model. *J. Adv. Manuf. Process* **2022**, *4*, No. e10118.

(19) Ismail, S.; Abdulkadir, R. A.; Usman, A. G.; Abba, S. I. Development of chemometrics - based neurocomputing paradigm for simulation of manganese extraction using solid - phase tea waste. *Model. Earth Syst. Environ.* **2022**, *8*, 5031−5040, DOI: 10.1007/s40808-022-01369-8.

(20) Usman, A. G.; IŞIK, S.; Abba, S. I. Qualitative prediction of Thymoquinone in the high-performance liquid chromatography optimization method development using artificial intelligence models coupled with ensemble machine learning. *Sep. Sci. Plus* **2022**, *5*, 579−587.

(21) Usman, A. G.; Işik, S.; Abba, S. I. Hybrid data-intelligence algorithms for the simulation of thymoquinone in HPLC method development. *J. Iran. Chem. Soc.* **2021**, *18*, 1537.

(22) Yang, L.; Liu, S.; Tsoka, S.; Papageorgiou, L. G. A regression tree approach using mathematical programming. *Expert Syst. Appl.* **2017**, *78*, 347−357.

(23) Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; de Marinis, G. Machine learning algorithms for the forecasting of wastewater quality indicators. *Water* **2017**, *9*, 105.

(24) Sampa, M. B.; Hossain, M. N.; Hoque, M. R.; Islam, R.; Yokota, F.; Nishikitani, M.; Ahmed, A. Blood uric acid prediction with machine learning: Model development and performance comparison. *JMIR Med. Inform.* **2020**, *8* (10), No. e18331, DOI: 10.2196/18331.

(25) Bhattarai, A.; Dhakal, S.; Gautam, Y.; Bhattarai, R. Prediction of nitrate and phosphorus concentrations using machine learning algorithms in watersheds with different landuse. *Water* **2021**, *13*, 3096.

(26) Wiangkham, A.; Ariyarit, A.; Aengchuan, P. Prediction of the influence of loading rate and sugarcane leaves concentration on fracture toughness of sugarcane leaves and epoxy composite using artificial intelligence. *Theor. Appl. Fract. Mech.* **2022**, *117*, No. 103188.

(27) Abba, S. I.; Benaafi, M.; Usman, A. G.; Aljundi, I. H. Sandstone groundwater salinization modelling using physicochemical variables in Southern Saudi Arabia: Application of novel data intelligent algorithms. *Ain Shams Eng. J.* **2023**, *14*, No. 101894.

(28) Rascon, C.; Meza, I. Localization of sound sources in robotics : A review. *Rob. Auton. Syst.* **2017**, *96*, 184−210.

(29) Abba, S. I.; Usman, A. G.; IŞIK, S. Simulation for response surface in the HPLC optimization method development using artificial intelligence models: A data-driven approach. *Chemom. Intell. Lab. Syst.* **2020**;*201*. 104007.

(30) Kouadri, S.; Elbeltagi, A.; Islam, A. R. M. T.; Kateb, S. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Appl. Water Sci.* **2021**, *11*, No. 124974, DOI: 10.1007/s13201-021-01528-9.

(31) Shams, S. R.; Jahani, A.; Kalantary, S.; Moeinaddini, M.; Khorasani, N. Artificial intelligence accuracy assessment in NO2 concentration forecasting of metropolises air. *Sci. Rep.* **2021**, *11*, No. 1805.

(32) Uzun Ozsahin, D. U.; Precious Onakpojeruo, E. P.; Bartholomew Duwa, B.; Usman, A. G.; Isah Abba, S. I.; Uzun, B. COVID-19 Prediction Using Black-Box Based Pearson Correlation Approach. *Diagnostics* **2023**, *13*, 1264.

(33) Nourani, V.; Elkiran, G.; Abba, S. I. Wastewater treatment plant performance analysis using artificial intelligence - An ensemble approach. *Water Sci. Technol.* **2018**, *78*, 2064−2076.

(34) Demirci, M.; Üneş, F.; Körlü, S. Modeling of groundwater level using artificial intelligence techniques: A case study of Reyhanli region in Turkey. *Appl. Ecol. Environ. Res.* **2019**, *17*, 2651−2663.

(35) Mahmoud, K.; Bebiş, H.; Usman, A. G.; Salihu, A. N.; Gaya, M. S.; Dalhat, U. F.; et al. Prediction of the effects of environmental factors towards COVID-19 outbreak using AI-based models. *IAES Int. J. Artif. Intell.* **2021**, *10*, 35−42.

(36) Nasr, M. S.; Moustafa, M. A. E.; Seif, H. A. E.; El Kobrosy, G. Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment. *Alexandria Eng. J.* **2012**, *51*, 37−43.

(37) Elkiran, G.; Nourani, V.; Abba, S. I. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *J. Hydrol.* **2019**, *577*, No. 123962.

(38) Wang, X.; Kvaal, K.; Ratnaweera, H. Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. *J. Process Control* **2019**, *77*, 1−6.

(39) Marrero-Ponce, Y.; Barigye, S. J.; Jorge-Rodríguez, M. E.; Tran-Thi-Thu, T. QSRR prediction of gas chromatography retention indices of essential oil components. *Chem. Pap.* **2018**, *72*, 57−69.

(40) Frank, L. E.; Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109−135.

(41) Abba, S. I.; Benaafi, M.; Usman, A. G.; Aljundi, I. H. Inverse groundwater salinization modeling in a sandstone's aquifer using stand-alone models with an improved non-linear ensemble machine learning technique. *J. King Saud Univ., Comput. Inf. Sci.* **2022**, *34*, 8162.

(42) Kawai, E.; Sato, H.; Furuichi, K.; Toru, T.; Yoshioka, T. Maximizing margins and optimizing operational conditions for residue fluid catalytic cracking with an artificial intelligence hybrid reaction model. *J. Adv. Manuf. Process* **2022**, *4* (3), No. e10118.

(43) Yang, F.; Dai, C.; Tang, J.; Xuan, J.; Cao, J. A hybrid deep learning and mechanistic kinetics model for the prediction of fluid catalytic cracking performance. *Chem. Eng. Res. Des.* **2020**, *155*, 202−210.

(44) Zhu, W.; Liu, X.; Hou, X.; Hu, J.; Diao, Z. Application of machine learning to process simulation of n-pentane cracking to produce ethylene and propene. *Chin. J. Chem. Eng.* **2020**, *28*, 1832−1839.

(45) Zhu, X.; Wang, N. Splicing process inspired cuckoo search algorithm based ENNs for modeling FCCU reactor-regenerator system. *Chem. Eng. J.* **2018**, *354*, 1018−1031.