http://dx.doi.org/10.12785/ugric/2025008

Developing a Real-Time Soft Sensor for Product Composition Estimation

Bassam Mohamed Alhamad
Department of Chemical Engineering
University of Bahrain
Isa Town, Bahrain
balhamad@uob.edu.bh

Rim Ahmed AlGendi
Department of Chemical Engineering
University of Bahrain
Isa Town, Bahrain
reemalgendy1@gmail.com

Abstract— **With a main objective of enhancing process stability and control, this paper explores the development and use of a real-time soft sensor for predicting product composition in Crude Distillation Units (CDUs). Developing a machine-learning model able to continually analyze and forecast product composition during crude oil distillation takes front stage. The basis of the research is experimental data derived from dynamic simulations of the CDU process employing Aspen-HYSYS form. Time Series Linear Regression (TSLR), Time Series Partial Least Squares (TSPLS), and Time Series Neural Networks (TSNN) are among the approaches used in several soft sensor models created. Performance measures, including root mean square (RMS) and the coefficient of determination (R-squared), guide evaluation of these models. Thanks to its better accuracy and predictive capabilities, where it obtained the lowest Root Mean Square (RMS) error of 0.8006 and the highest coefficient of determination (R-squared), the Time-series Neural Network (TSNN) stands out among the developed models as the best option for distillation endpoint estimate in CDUs. Then linked into the Aspen HYSYS modelling plant, the TSNN soft sensor estimated diesel molar flow in real-time. Integrated with the simulated plant, the model was trained on real-time data originating from an Aspen HYSYS simulation of a crude oil distillation unit, allowing continuous live estimates.**

Keywords— Crude Distillation Units (CDUs), soft sensors, Time Series Neural Networks (TSNN), Aspen HYSYS simulation, product composition prediction, Root Mean Square (RMS) error, and machine learning models.

I. Introduction

Within the field of industrial process optimization, especially in relation to crude oil refining, the search for improved efficiency and quality control still presents a great difficulty. The first refining process depends much on crude distillation units (CDUs), which separate crude oil into many fractions depending on boiling points. Ensuring optimal performance and satisfying strict product quality criteria depend on accurate estimates of product composition throughout CDU operations.

Current approaches monitor and control product compositions in CDUs mostly using physical sensors and laboratory analysis. Although these conventional methods offer insightful analysis, they are sometimes constrained by high prices, maintenance needs, and difficulty to record real-time changes in product compositions. This dependence on offline measurements emphasizes the requirement of more flexible and affordable technologies able to run perfectly inside the dynamic surroundings of a refinery.

The main goal of this work is to provide solutions by means of a real-time soft sensor meant to estimate CDU product compositions. Writing the mathematical model is a tedious work that could be replaced by developing a soft sensor that is trained through machine learning. The learning process of the soft sensor relies on the accuracy of the model that is developed to replicate as a digital twin with the industrial plant. These soft sensors would provide the CDU product composition for monitoring and control. To provide a better learning process, historical data is used across different time zones that are used to validate the developed model. While training, filtering information and spotting any anomalies is critical to having an accurate soft sensor. Methodologies will be used to simplify the process of soft sensing.

Soft sensors have been created in particular contexts of predicting the product composition by means of continuous readings from process streams in a refinery crude distillation unit. Partial least squares, artificial neural networks, and linear regression analysis have been applied in the construction of both multiple linear and nonlinear soft sensor models.

A. Aim of the paper

The objective of the project is to create a dependable computational model or algorithm capable of accurately forecasting the distillation endpoint in real-time during the crude oil distillation process in a CDU, employing a soft sensor that utilizes advanced mathematical or artificial intelligence techniques to analyze extensive process data and generate precise predictions of the distillation endpoint. When executed proficiently, this technology may significantly boost the operating efficiency of the crude distillation unit, improve product quality, and optimize process control.

B. Motivation

For the purpose of determining the distillation endpoint in crude distillation units, traditional monitoring approaches are not only expensive but also less reliable. As a result, the refining sector is now required to optimize efficiency, improve process control, and comply with specifications and regulations. Through the use of soft sensors, this work will deliver a reliable and cost-effective real-time estimating solution, therefore reducing the dependency on traditional monitoring approaches and enhancing the stability and controllability of refining operations. This will assist to boost productivity in CDUs, as well as product quality and compliance with regulatory requirements.

II. Literature Review

The conversion of crude oil into products such as diesel, kerosene, and naphtha cover the foundation of the energy sector. Central to this process are Crude Distillation Units (CDUs), which necessitate precise control systems to uphold efficiency and guarantee product quality. Traditionally, physical sensors have been employed for measurement and

process management in these systems. Nonetheless, they include disadvantages, such as substantial expenses for installation and upkeep, in addition to delays stemming from dependence on manual sampling and sluggish analytics.

The use of soft sensors has garnered much interest to address these constraints. These data-driven models predict process variables using accessible inputs, offering quicker and more economical options compared to conventional sensors [1], [2]. Recent advancements in machine learning (ML) and artificial neural networks (ANNs) have significantly enhanced the functionality of soft sensors, enabling real-time monitoring of intricate systems [3], [4], [5]. This paper examines improvements, practical applications, obstacles, and prospects for the integration of soft sensors with advanced technologies like digital twins.

Soft sensors have progressed markedly in conjunction with improvements in computing techniques and algorithms. Initial versions, as emphasized by [1], were based on linear and nonlinear modeling methodologies. These models relied significantly on data preparation, including outlier detection and dataset normalization, to enhance their prediction efficacy. As system complexity increased, linear models failed to adequately represent the complicated dynamics of CDU operations, necessitating the use of increasingly sophisticated nonlinear methodologies.

Kubosawa et al. (2022) demonstrated the potential of hybrid models that integrate dynamic simulations with artificial intelligence techniques [3]. The integration of approaches not only augmented the flexibility of soft sensors to fluctuating operating circumstances but also fortified their resilience. Lüthje et al. (2020) illustrated the use of hybrid models, which combine data-driven approaches with mechanical process comprehension, to nonlinear predictive control, yielding dependable results across diverse situations [6].

Artificial neural networks are very adept at managing the nonlinearities intrinsic to CDU systems. By analyzing previous data, these networks can forecast essential factors, like feed composition and the quality of final goods. GaJang et al. (2010) demonstrated the application of feedforward neural networks in mapping intricate interactions between input and output variables [7].

Kataria and Singh (2017) achieved more progress by utilizing recurrent neural networks (RNNs) for temporal data. The capacity of RNNs to handle sequential data via feedback loops renders them very suitable for dynamic processes such as crude oil distillation. Their research emphasized the enhanced efficacy of RNN-based sensors in environments with fluctuating operating parameters, in contrast to conventional static models [8].

Expanding upon this foundation, Park et al. (2015) created ANN-based soft sensors for real-time feed monitoring, therefore diminishing the necessity for physical analyzers [2], [9]. The capacity to adjust to real-world data and deliver actionable insights highlights the significance of ANNs in contemporary refining operations [7], [10], [11].

In CDU situations, where conditions are always changing, the analysis of time-series data is essential. Advanced designs, like Long Short-Term Memory (LSTM) networks, have demonstrated notable efficacy for certain tasks. Chatterjee and Saraf (2004) investigated the use of LSTM models for forecasting product compositions in distillation processes, notwithstanding the presence of noise or incomplete data. The selective memory retention properties of LSTMs guarantee precision and dependability in these dynamic systems [12].

A primary advantage of soft sensors is their capacity to deliver real-time insights on process factors. Oster et al. (2023) illustrated the application of machine learning-driven soft sensors in vacuum distillation for the continuous and precise forecasting of product characteristics [4]. This capacity allows operators to expedite informed decision-making, hence improving overall system efficiency and product quality.

Soft sensors are essential for predictive maintenance. These systems mitigate unexpected downtime by analyzing patterns and identifying early warning indicators of equipment deterioration. Barbosa (2014) created a soft sensor to assess the quality of hydrocracker products, which facilitated process management and offered critical insights into equipment health. This proactive strategy guarantees the optimization of maintenance schedules and the reduction of interruptions [13].

The efficacy of soft sensors is largely contingent upon the quality of the data utilized for their training. Inaccurate, deficient, or prejudiced datasets can profoundly impact model precision [14]. To resolve these challenges, approaches like noise reduction, feature engineering, and data augmentation must be employed during the preprocessing phase.

Although machine learning models exhibit excellent accuracy, they frequently face criticism for their insufficient interpretability. Operators in essential sectors, like as refining, must rely on the insights offered by these models. Hybrid methodologies, integrating domain knowledge with data-centric modeling, provide a more visible and comprehensible framework [15].

Dynamic operational contexts necessitate models capable of adapting to fluctuating input data instantaneously. Kim et al. (2022) investigated adaptive learning algorithms that may self-update, therefore obviating the necessity for regular manual retraining and maintaining consistent performance across diverse settings.

Digital twins—virtual representations of physical systems—are transforming process optimization. When combined with soft sensors, digital twins offer an extensive framework for monitoring and predictive analysis. Li and Sun (2023) showed that the integration of these technologies improves decision-making and process efficiency in refinery operations, facilitating the development of more intelligent and responsive systems [15].

Soft sensors, driven by advancements in artificial intelligence and machine learning, signify a significant advancement in CDU process management [9], [16]. They provide real-time surveillance, anticipatory maintenance, and improved operational efficacy. Confronting issues like data integrity, openness, and flexibility will be essential for optimizing their potential. The amalgamation of soft sensors with digital twins underscores their revolutionary influence, solidifying their status as a fundamental element of innovation in the refining sector.

III. Design and Implementation

This study utilizes HYSYS's dynamic simulation model, which is already under control, to generate data for understanding the dynamic behavior of a Crude Distillation Unit (CDU) with respect to its molar flow of diesel. Inspired by the research in "Development of Soft Sensors for Crude Distillation Unit Control" by Mohler et al., we focus on six key input variables identified as the most likely to influence diesel output: column top temperature, column bottom temperature, light gas oil temperature, heavy gas oil temperature, pump-around temperature, and pump-around flow rate.

Each input variable is controlled by a PID controller within the HYSYS model. By implementing step changes in the setpoints of these PID controllers, we simulate disturbances and observe the resulting dynamic response of the CDU's molar flow of diesel. This approach generates valuable time-series data, capturing how the CDU behaves under varying operating conditions triggered by changes in these critical input variables. Analyzing this data will be instrumental in gaining insights into the dynamic relationship between these key factors and the molar flow of diesel, paving the way for potential future control strategy improvements.

As the first step, I selected relevant data from the plant database, which was obtained through dynamic simulations in Aspen HYSYS. There are six input variables chosen to analyze their effect on the diesel composition over time: column top temperature, kerosene temperature, diesel temperature, diesel pump-around temperature, diesel pump-around flow rate, and column bottom temperature. The diesel molar flow (composition) served as the output variable. To investigate these relationships, seven PID controllers were implemented in HYSYS. Six controllers manipulated the input variables, while the final one regulated the diesel molar flow, as shown in Fig. 1. The simulation of the whole plant is shown in Fig. 2.

The first PID controller, FIC-102, is developed to maintain the desired molar flow rate of the diesel product. On the Connections tab in Fi, the process variable (PV) was defined in the first. This variable represents the actual diesel molar flow measured by the plant, which FIC102 will continuously monitor and attempt to regulate at the setpoint. Second, the output target object (OT) was designated. This signifies the element the controller will adjust to influence the process variable. In this case, FIC-102 manipulates the actuator position of valve VLV-100 to control the diesel molar flow.

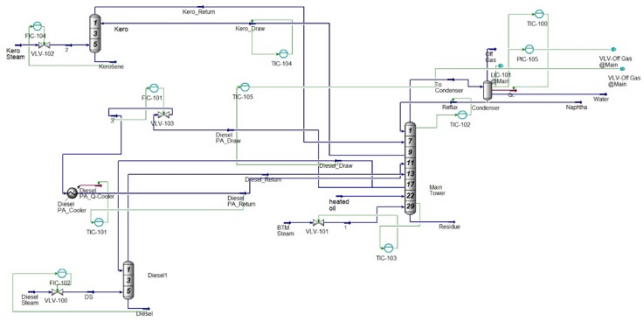

Fig. 1. The simulate process flow diagram of the CDU plant

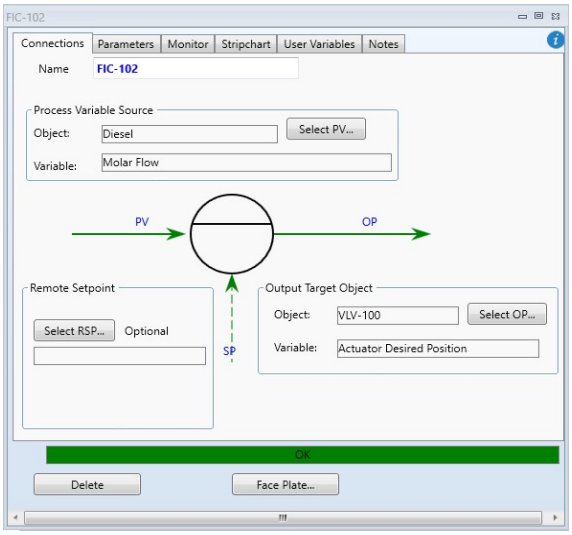

Fig. 2. Connection's tab of FIC-102 controller

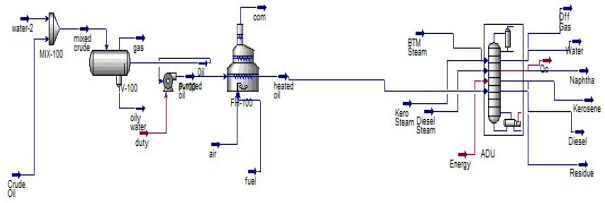

Fig. 3. Simulation of the whole plant

Next, on the Parameters tab, as in Fig. 4, operational parameters for FIC-102 were specified. The action mode was set to "Direct", indicating a proportional relationship between the controller output and the valve position. Additionally, the setpoint, the desired diesel molar flow value, was defined along with the minimum and maximum acceptable values for the process variable. Finally, tuning parameters, which influence the controller's responsiveness and stability, are established based on process knowledge and engineering judgment.

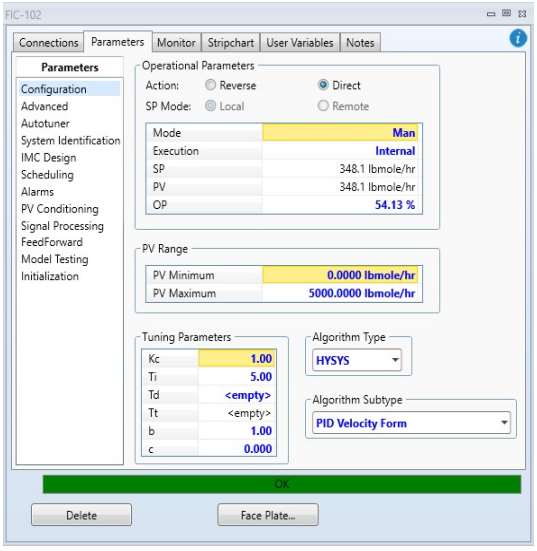

Fig. 4. Parameter's tab of FIC-102 controller

Once configured, the controller was activated by switching its mode to "Auto" or "Man". The plant is initially operated in manual mode to observe the impact of the input variables on the diesel output. However, to gather data for analysis, model testing was conducted using FIC-102. This testing process involved defining several parameters. A step change is introduced to have variations in the setpoint. The signal variation amplitude was determined based on the magnitude of this variation. The time interval was specified to produce the wanted frequency of the data points to be produced. The testing time length is defined to identify the total duration of the model test.

After setting these parameters, the simulation was run for six hours. The test results were exported to Excel. This data is used to build and train the soft sensor model. The plot of the data of the bottom temperature in a time series plot is shown in Fig. 6. In addition, the time series plot of the Diesel Molar Flow is also shown in Fig. 6. As there is some instability in the first readings produced, this will not be used in the training process, as it can potentially cause problems with the model if included in the training data. Hence, this data is removed to avoid building an inefficient model.

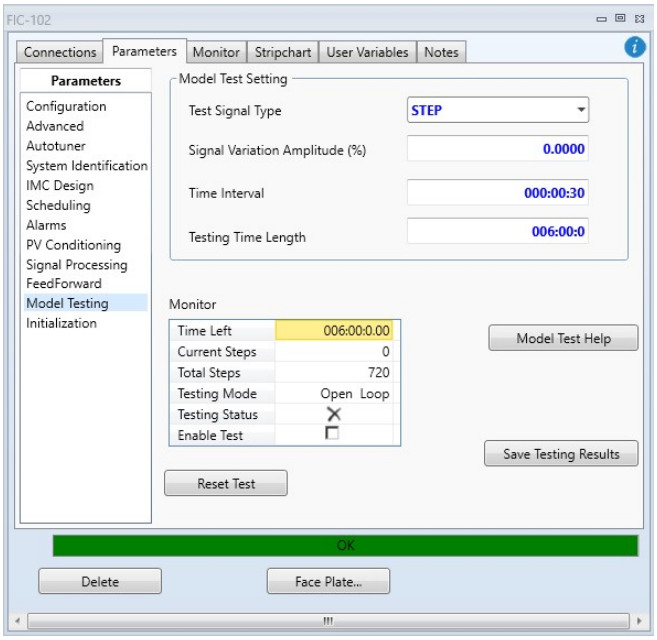

Fig. 5. Model testing tab of FIC-102 controller

A. *Data Analysis*

A correlation matrix is a fundamental statistical tool that provides insights into the relationships between variables in a dataset. It consists of a square matrix where each cell represents the correlation coefficient between two variables. Correlation coefficients measure the strength and direction of the linear association between variables, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, 0 represents no correlation, and -1 indicates a perfect negative correlation, whereas the red colors denote a positive correlation, and the blue represents a negative correlation shown in Fig. 8.

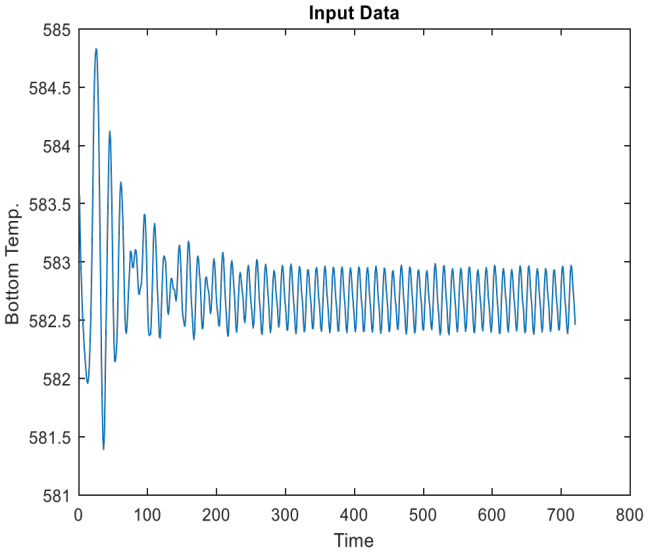

Fig. 6. Time series plot of the bottom temperature

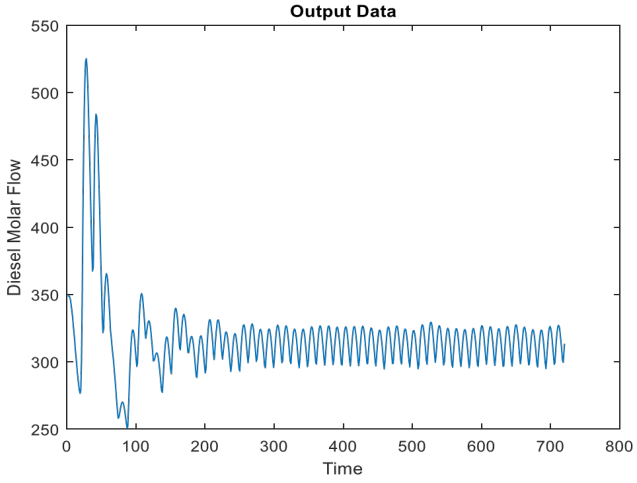

Fig. 7. Time Series Plot of the Diesel molar flow rate

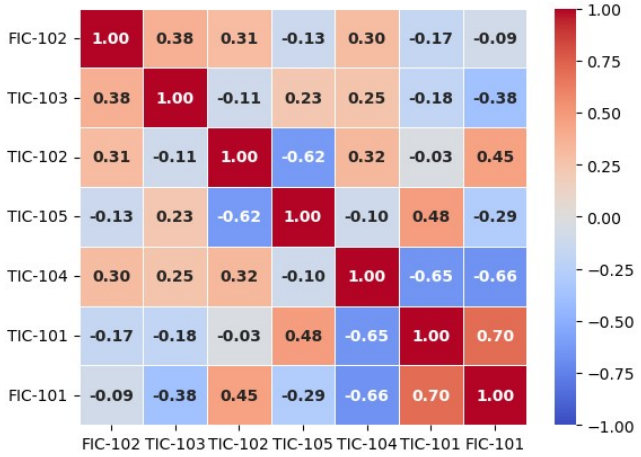

Fig. 8. Heat map of the correlation matrix

By examining the correlation matrix, researchers and analysts can identify patterns, determine the strength of relationships, and understand the dependencies between variables. This information is crucial for various data analysis tasks, such as identifying key factors, exploring multicollinearity, selecting variables for models, and gaining insights into the dataset's underlying structure. The correlation matrix is a powerful tool that aids in summarizing and visualizing the relationships between variables, providing a foundation for further analysis and interpretation.

The correlation matrix provided shows the correlation coefficients between the output variable output diesel molar flow and the input variables column top temperature, kerosene temperature, diesel temperature, diesel pump-around temperature, diesel pump-around flow rate, and column bottom temperature. The correlation coefficient between FIC-102 and itself shows the value of 1, as it represents the correlation of the variable with itself, which is perfect positive correlation.

Looking at the correlation coefficients, it is observed that diesel molar flow (FIC-102) has a weak positive correlation with TIC-102 (0.3063), while TIC-103 (0.3784) has the strongest correlation with the output, indicating that as the input variable increases, FIC-102 tends to slightly increase. On the other hand, FIC-102 has a weak negative correlation with TIC-105 (0.1284), TIC-101 (-0.1669), and FIC-101 (-0.0914), implying that as these variables increase, FIC-102 tends to decrease slightly.

There are some correlations between the input variables that are worth noting. As we can see,

TIC-102 and TIC-105 exhibit a strong negative correlation (-0.6196), while TIC-104 and TIC101 show a strong negative correlation (-0.6514). These high-magnitude correlations suggest the presence of collinearity between these pairs of variables. The cross-relation of bottom temperature and diesel molar flow is shown in Fig. 9.

Cross-correlation analysis is also a valuable tool for investigating the relationship between two time series data sets, particularly in how changes in one variable might influence another with a possible time lag. In the simulated case, the cross-correlation is used to analyze the relationship between the manipulated operating variables (such as column top temperature) and the diesel molar flow throughout the simulation. This will help in identifying how adjustments to these operating conditions might impact the diesel production over time. By observing the shift (time lag) at which the correlation between variables peaks, one can gain insights into the cause-and-effect relationships within the process.

In the relationship between two time series (x(t) and y(t)), the series may be related to past lags of the x-series. The sample cross-correlation function (CCF) is helpful for identifying lags of the x-variable that might be useful predictors of y(t).

x-variable(s) will be set to be the leading variable of the y-variable to predict future values of y. Thus, one will usually be looking at what's happening at the negative values of the cross-correlation plot as shown in Fig. 8.

In Fig. 10, the cross-correlation (CC) plot will be used to determine the time lags that have strong correlations with future y values, and these lagged x variables will be included with the input data to build the machine learning models. However, adding too much data will further complicate the model, so that's why only the three most significant correlations are going to be included.

B. Model Development

The data is initially generated in HYSYS and subsequently employed in the development of three distinct varieties of time series models in MATLAB. The procedure commences with two linear models, TSMLR and TSPLS, and culminates with the nonlinear TSNN model. The results of each model are analyzed to ascertain which one possesses the most predictive capacity. Consequently, these models are implemented to forecast the composition of diesel.

Commencing with the TSMLR model, the lagged variant of the original data was implemented to improve the precision of our model. Our investigation subsequently included the k-1 lag, k-2 lag, and k-3 lag as predictors, which were identified through cross-correlation. The data was partitioned into a training set, which contains 70% of the original dataset for coefficient estimation, and a testing set, which contains the remaining 30% to assess the model's performance in order to ensure the model's resilience. This allocation enables us to assess the efficacy of our model on data that is both novel and previously unexplored. The "regress()" function in MATLAB is employed to develop the prediction model, which is based on sophisticated regression methodologies.

Related to TSPLS reflecting the time series modeling, lagged duplicates of the original data were implemented as predictors. To identify potential temporal correlations in the data, cross-correlation analysis was implemented by selecting delays of k-1, k-2, and k-3. The data was subsequently divided into training (70%) and assessment (30%) sets. The relationship between the input variables and the objective variable was modeled using partial least squares (PLS) regression using the "plsregress" function in MATLAB. Twenty-four latent variables were chosen within the PLS framework to accurately represent the fundamental structure of the data.

The TSNN architecture was generated using the "ntstool" program in MATLAB. An iterative approach is employed to determine the optimal network architecture. A fundamental architecture is developed with a single concealed layer and six neurons, which is equivalent to the number of input variables. Consequently, the predictive efficacy was improved by increasing the number of neurons. The final network consisted of a single concealed layer that contained ten neurons. This methodology incorporated a three-second time delay as a conspicuous characteristic. This enabled the network to encode temporal relationships by learning from both the current input and historical values in the time series. The concealed layer implemented the "tanh" activation function to adapt to modeling non-linear correlations in the data. To generate continuous output values that are pertinent to regression tasks, the linear activation function was employed in the output layer. The "Levenberg-Marquardt" method, which is widely recognized for its efficacy, was employed to train the network. To mitigate overfitting, a deliberate data partitioning strategy was implemented: 70% of the data was allocated for training, 10% for validation during the training process to assess performance, and the remaining 20% was reserved for final testing and assessment.

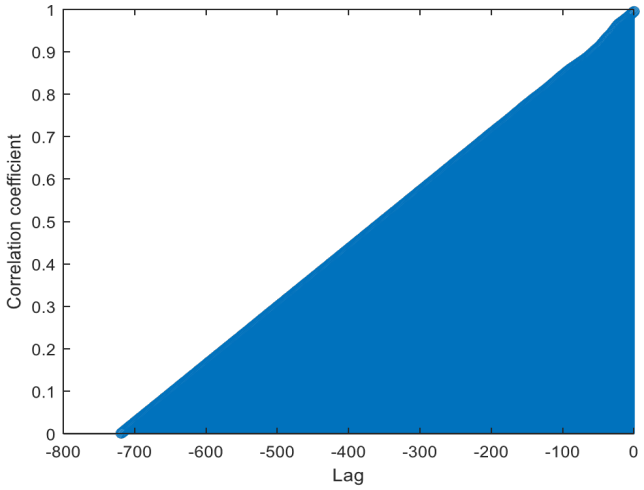

Fig. 9. Cross-correlation of bottom temperature TIC-103 and diesel molar flow

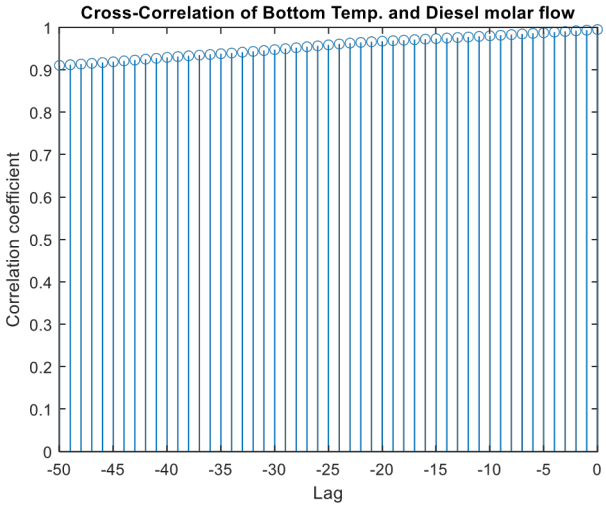

Fig. 10. CC plot of bottom temperature and diesel molar flow

The objective of controlling the fractionator is to maintain the desired values for the endpoints of the top-draw product (y1), side-draw product (y2), and bottom reflux temperature (y7). This is achieved by manipulating the flow rates of the top draw (u1), side draw (u2), and the heat transfer rate of the bottom reflux (u3). The heat transfer rate (u3) is further adjusted using a control loop that utilizes the hot steam flow rate as a control variable. Additionally, there are two measured disturbances in the system: the heat transfer rate of the upper reflux (l1) and the intermediate reflux (l2). These flows remove heat from the system and are subsequently reboiled in other sections of the plant.

TABLE I. STATISTICAL PARAMETERS OF TSMLR MODEL

Parameters			
R^2	*F-test*	*p-value*	*RMSE*
0.7058	6.75	0.0000	11.2509

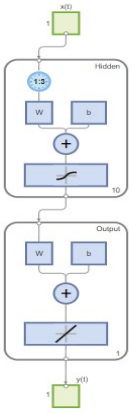

Fig. 11. TSNN architecture

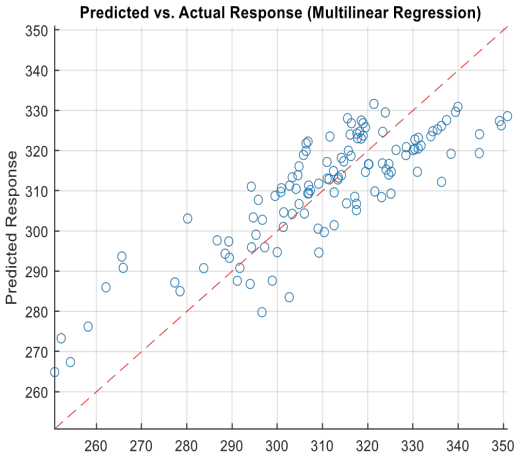

Fig. 12. Scatter plot to TSMLR model

IV. RESULTS AND DISCUSSIONS

A. *TSMLR model development results*

Based on the analysis of the linear model, Table I presents the statistical parameters. The coefficient of determination (R^2) value of 0.7058 indicates a relatively good level of explained variability in the response variable by the predictors. Although not exceptionally high, it suggests that the model captures a significant portion of the data's variation. The F-test yielded a significant F-value of 6.75 (p-value = 0.0000), indicating a rejection of the null hypothesis of no linear dependence; in other words, there is no autocorrelation of error relation. This confirms the presence of a relationship between the predictors and the response variable. It is worth noting that the Root Mean Squared Error (RMSE) of 11.2509 represents the average deviation of the model's predictions from the actual values, suggesting an acceptable level of accuracy. Overall, while the model performs well, there is room for improvement in terms of capturing more of the variability and reducing prediction errors.

The time series plot shows in Fig. 12 and Fig. 13 that the model's predictions (orange line) generally follow the trend of the actual values (blue line). However, there are some deviations between the two lines. This indicates that the model's predictions are not perfectly accurate, but they are on the right track.

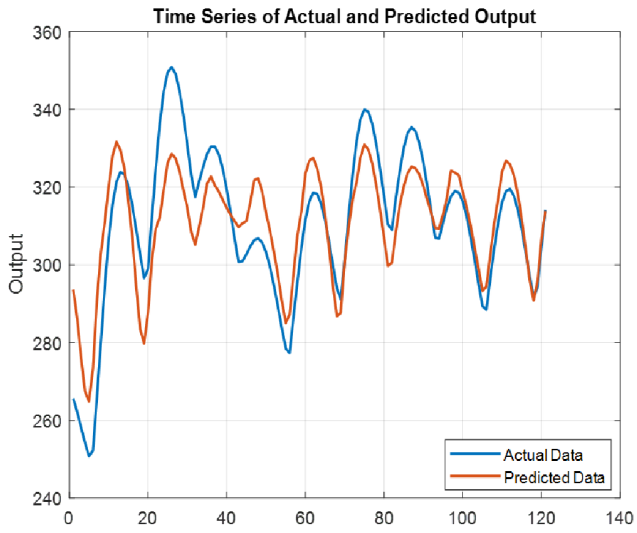

Fig. 13. TSMLR model predications vs. actual data

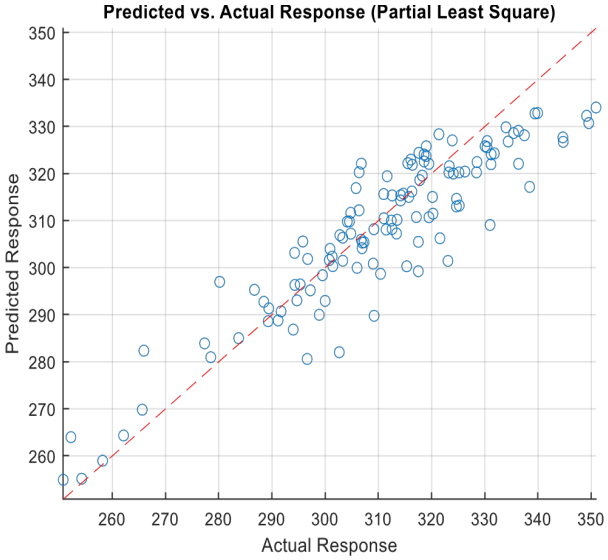

Fig. 14. Scatter plot to TSPLS model

B. *TSPLS model development results*

The Time Series Partial Least Square (TSPLS) model yielded highly promising results, as indicated by the statistical parameters presented in Table II. The correlation coefficient (R) between the predicted and actual values was found to be 0.90993, suggesting a strong positive linear relationship. This implies that the TSPLS model effectively captured the underlying patterns and trends present in the time series data. The coefficient of determination (R-squared or R^2) was calculated to be 0.82797, indicating that approximately 82.8% of the variance in the response variable could be explained by the predictors included in the model. This demonstrates the model's ability to effectively explain and predict the observed outcomes. The Root Mean Squared Error (RMSE) was determined to be 8.7786, which represents the average deviation between the predicted and actual values. A lower RMSE value signifies a higher level of accuracy, suggesting that the TSPLS model exhibited satisfactory predictive performance. Additionally, an F-test was conducted, resulting in an F-value of 7.2193 and a corresponding p-value of $6.5281e^{-14}$. The significant F-test indicates that the predictors included in the model collectively have a significant impact on predicting the response variable. In summary, the TSPLS model demonstrated strong predictive power, capturing the underlying patterns in the time series data with high accuracy.

From Fig. 14 and Fig. 15, it can be observed that the predicted values closely align with the actual values. This alignment indicates that the TSPLS model has a good level of accuracy in forecasting and analyzing the time series data. The proximity of the actual and predicted lines suggests that the model effectively captures the trends and fluctuations in the data.

TABLE II. STATISTICAL PARAMETERS OF TSPLS MODEL

Parameters			
R^2	***F-test***	***p-value***	***RMSE***
0.82797	7.2193	$6.5281e^{-14}$	8.7786

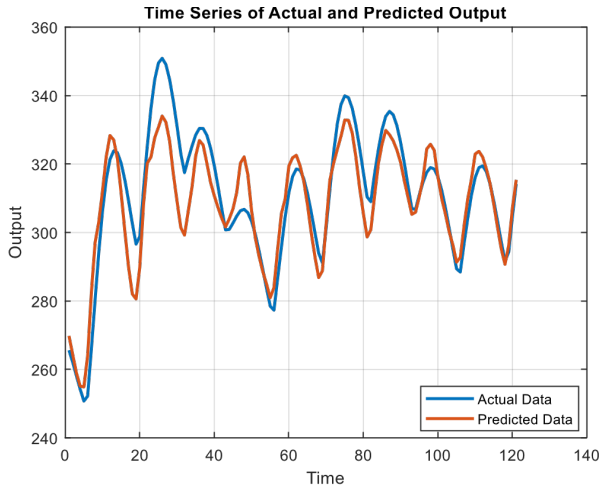

Fig. 15. TSPLS model predictions vs. actual data

C. *TSNN model development results*

The Time Series Neural Network (TSNN) emerged as the top performer among the models evaluated, achieving the lowest Root Mean Squared Error (RMSE) and highest R-squared values across all datasets. As summarized in Table III, it achieved the lowest Root Mean Squared Error (RMSE) of 0.1659 during training, indicating high accuracy in estimating product composition in the Crude Distillation Unit (CDU). Furthermore, TSNN exhibited impressive correlation coefficients, with an R-value of 0.9995 during training and an R^2 value of 0.9997, indicating a strong linear relationship between the predicted and actual values. These results were further validated during the validation and test stages, where TSNN achieved RMSE values of 0.5245 and 0.8008, respectively, as well as high R and R-squared values of 0.9982, 0.9991 and 0.9968, 0.9984, respectively.

Further analysis of the results are generated as shown in Fig. 16 and Fig. 17. Fig. 16 illustrates the histogram of target-output errors with 20 bins. Analyzing the plot, this study notes that errors fall within a very small value of 0.007254. All the error training points are close to this point. Consequently, the error histogram in this context emphasizes extremely good outcomes.

TABLE III. STATISTICAL PARAMETERS OF TSNN MODEL

Testing	Parameters		
	RMSE	*R*	R^2
Training	0.1659	0.9995	0.9997
Validation	0.5245	0.9982	0.9991
Test	0.8006	0.9968	0.9984

The performance plotted in Fig. 17 shows iterations that the process of training has been done. It specifies how much the final error and gradient are. The training process ended at the 53rd iteration, where it is not the best choice for a trained network. The algorithm chooses the 47th iteration because it has a less valid error in comparison to the training error. It means that, by continuing the process, the iteration may have a performance for training data, but this is not done to prevent overfitting causing lower performance. The minimum mean squared error (MSE) is 0.52451 at epoch 74.

This study uses the error autocorrelation function to examine how the predicting errors are interconnected in time to verify the network's performance. In this case, Fig. 18 reveals a single nonzero value at zero lag, representing the mean square error, which is approximately 0.85, indicating a high level of accuracy. Apart from the zero-lag correlation, most of the predicting errors fit inside the confidence limit around zero, confirming the efficiency of the predicting approach.

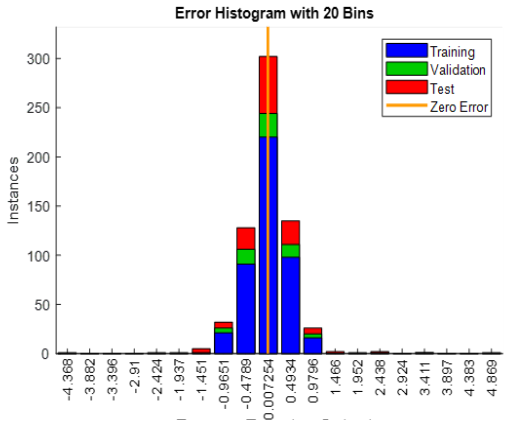

Fig. 16. Error histogram with 20 bins

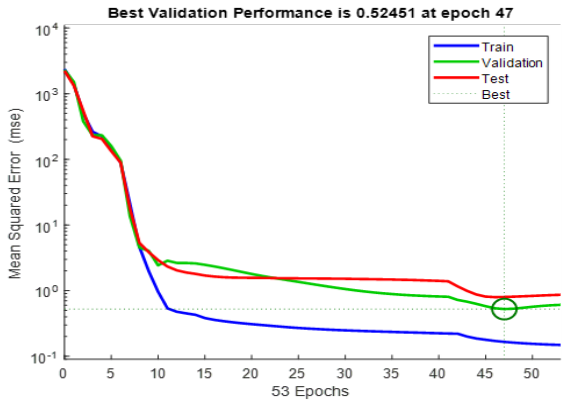

Fig. 17. The best performance of training, testing, and validation using TSNN

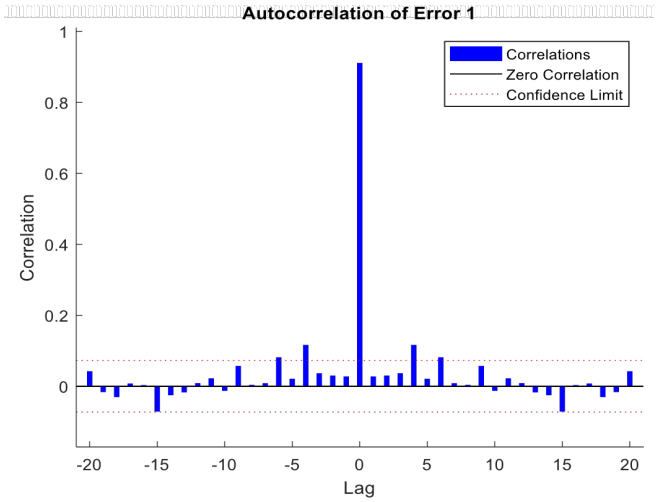

Fig. 18. Autocorrelation of error

Fig. 19 depicts the model's performance in predicting diesel output. A visual comparison between the model's predictions (outputs) and the actual measured diesel production data (targets) is presented as a time series plot. This plot allows us to observe the response of the neural network, comparing the predicted values with the actual values over time. Additionally, a plot of the prediction errors (the difference between predicted and actual values) is included. The model successfully captures the overall trend of diesel production, demonstrating its ability to represent the target behavior. Furthermore, most of the prediction errors fall within a range of less than -5 to 5, indicating a high accuracy.

The time series neural network continued to showcase its exceptional performance, achieving impressive outcomes against additional tests. Table IV presents the results obtained, where it achieved an RMSE value of 0.1933, indicating a high level of accuracy in estimating product composition. Moreover, a high R-squared value of 0.9995 is achieved, reflecting the accuracy and reliability of TSNN in predicting product composition. These findings reinforce the robustness and effectiveness of the TSNN model in real-world scenarios and solidify its position as the top performer among the evaluated models for arterial intelligence.

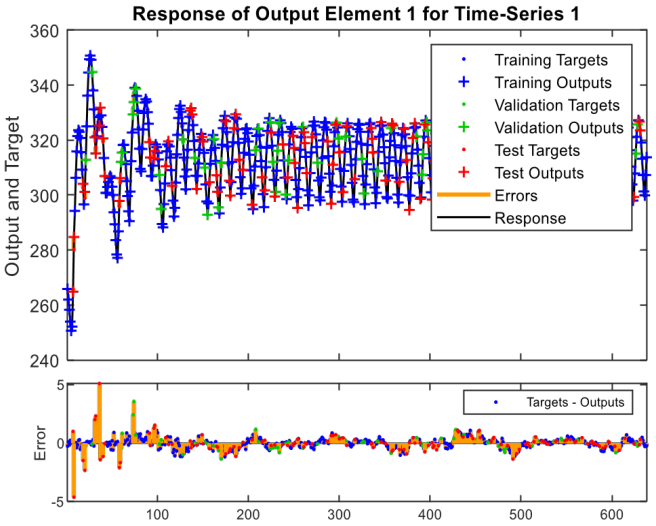

Fig. 19. Time series response plot of target series

TABLE IV. ADDITIONAL TEST STATISTICAL PARAMETERS OF TSNN MODEL

Testing	Parameters		
	RMSE	***R***	R^2
Additional test	0.1933	0.9990	0.9995

D. Real-time soft sensing implementation

The implementation of the TSNN model as a real-time soft sensor in Aspen HYSYS was done using the three different models, TSMLR, TSPLS, and TSNN. By seamlessly integrating the model into the simulation environment, the continuous estimation of the properties of CDU products was enabled based on live measurements of key input variables. This integration was made possible through a custom MATLAB-HYSYS interface code, which acts as a communication bridge between the platforms.

To facilitate this integration, the comprehensive MATLAB-HYSYS interface code allows for real-time data exchange and interaction between the soft sensor model in MATLAB and the simulation plant in HYSYS, which is validated. With the integration in place, the soft sensor now receives real-time input data from HYSYS, enabling it to make accurate estimations of the diesel molar flow.

Fig. 20 demonstrates the prediction accuracy of the real-time soft sensor. where the predicted values closely match the actual values, indicating the model's exceptional accuracy in real-time forecasting. The near-perfect proximity of the lines suggests that the soft sensor effectively captures the trends and dynamic fluctuations observed in the direct measurement data.

Table V shows the RMSE and R values for the three different models. The performance of the soft sensors was best with the TSNN model, TSPLS, and TSMLR, respectively. The real-time soft sensing was achievable by all different methods; however, with variable accuracy relying on the developed model. This is in agreement with the visual results in Fig. 20.

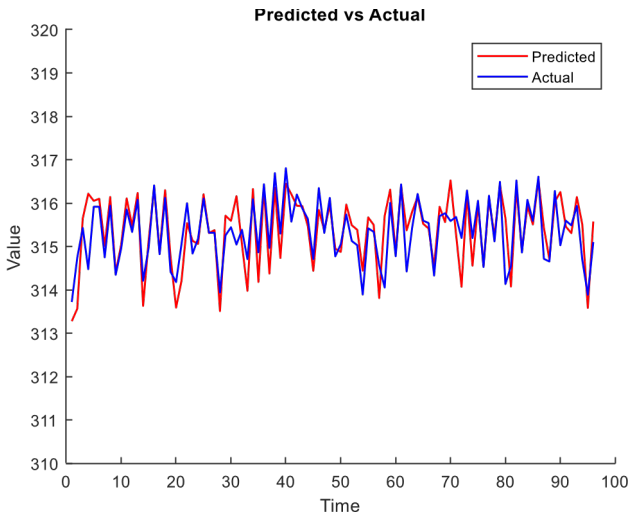

Fig. 20. Real-time soft sensor prediction based on live measurements of input variables.

TABLE V. PERFORMANCE COMPARISON OF SOFT SENSOR MODELS

Model	Parameters	
	RMSE	***R***
TSMLR	11.2509	0.7058
TSPLS	8.7786	0.82797
TSNN	0.8006	0.9968

V. CONCLUSIONS

The paper sought to enhance product composition monitoring in crude oil distillation units (CDUs) by the creation of a real-time soft sensor for estimating distillation endpoints. Conventional approaches are expensive, ineffective, and devoid of internet metrics. The study effectively developed a computational model to forecast distillation endpoints in real-time utilizing diverse machine learning methods. Dynamic process data was produced from an Aspen HYSYS simulation, thereafter preprocessed and analyzed to ascertain input variables and identify outliers or absent values.

Three soft sensor models were created employing distinct time-series methodologies: Time Series Multi Linear Regression (TSMLR), Time Series Partial Least Squares (TSPLS), and Time Series Neural Network (TSNN). A systematic escalation in model complexity was implemented, with each successive model enhancing the preceding one. This methodology facilitated a systematic assessment and comparison of model efficacy. The TSMLR model established a preliminary baseline but demonstrated restricted accuracy, with an RMSE of 11.2509. The TSPLS method enhanced this with an RMSE of 8.7786 by identifying nonlinear interactions via latent variables. The TSNN model yielded the most favorable outcomes, with an RMSE of 0.8006 and an R-squared value of 0.9968 during validation with previously unexamined test data. The optimized TSNN soft sensor was subsequently installed in the Aspen HYSYS environment to illustrate its capability for continuous live estimates of the distillation endpoint. The successful implementation of the soft sensor was achieved by providing continuous and reliable estimations of product properties, which would support monitoring, process control, stability enhancement, decision-making, and quality control.

REFERENCES

[1] A. Ž. Ujević Andrijić, I. Mohler, and N. Bolf, "Development of soft sensors for CDU control," Refin. Technol., vol. 3, no. 4, pp. 231–239, 2011.

[2] S. J. Park, "Neural network-based software sensors in CDUs," Refin. Syst., vol. 4, no. 5, pp. 77–84, 2015.

[3] S. Kubosawa, T. Onishi, and Y. Tsuruoka, "AI-enhanced soft sensors for chemical facilities," AIChE J., vol. 68, no. 7, pp. 44–56, 2022.

[4] J. Oster, M. Braun, L. Meier, and P. Schmidt, "ML-based soft sensors in vacuum distillation," Chem. Eng. Res. Des., vol. 179, no. 2, pp. 567–576, 2023.

[5] R. Shukla, P. Verma, M. Agarwal, and N. Gupta, "Dimensionality reduction in ML-based soft sensors," AI Chem. Eng., vol. 62, no. 3, pp. 89–102, 2020.

[6] T. Lüthje, M. Schmidt, P. Weber, and A. Krause, "Hybrid models in nonlinear predictive control," J. Process Control, vol. 85, no. 6, pp. 128–141, 2020.

[7] H. GaJang, K. Lee, M. Kim, and S. Park, "ANN-based feed identification in crude distillation," Comput. Chem. Eng., vol. 34, no. 5, pp. 731–742, 2010.

[8] G. Kataria and H. Singh, "Dynamic neural networks for reactive distillation," J. Process Control, vol. 53, no. 1, pp. 21–33, 2017.

[9] Y. S. Perera, D. A. A. C. Ratnaweera, C. H. Dasanayaka, and C. Abeykoon, "The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review," Eng. Appl. Artif. Intell., vol. 121, p. 105988, 2023, doi: https://doi.org/10.1016/j.engappai.2023.105988.

[10] S. Chatterjee and D. N. Saraf, "Artificial neural network models for dynamic systems," Chem. Eng. Sci., vol. 59, no. 3, pp. 637–649, 2004.

[11] L. T. Popoola, G. Babagana, and A. A. Susu, "Expert system design and control of crude oil distillation column of a Nigerian refinery using artificial neural network model," 2013. [Online]. Available: http://arpapress.com/volumes/vol15issue3/ijrras_15_3_16.pdf.

[12] D. Morris, P. Davis, R. Sanders, and K. Thompson, "Applications of LSTMs in process optimization," J. AI Refin., vol. 28, no. 4, pp. 89–102, 2019.

[13] J. M. Barbosa, "Development of soft sensors for hydrocracker product quality prediction," J. Process Control, vol. 24, no. 2, pp. 150–160, 2014.

[14] L. Wang and Y. Chen, "The role of data preprocessing in ML," AIChE J., vol. 62, no. 7, pp. 177–186, 2016.

[15] Y. Li and X. Sun, "Integrating digital twins with soft sensors in refineries," Comput. Chem. Eng., vol. 172, no. 1, pp. 105–113, 2023.

[16] L. Peterson, I. V. Gosea, P. Benner, and K. Sundmacher, "Digital twins in process engineering: An overview on computational and numerical methods," Comput. Chem. Eng., vol. 193, p. 108917, 2025, doi: https://doi.org/10.1016/j.compchemeng.2024.108917