Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Development of an Intelligent Soft Sensor Using Time-Series Neural Networks for Real-Time Composition Prediction in CDUs

Bassam Alhamad[1], Rim Algendi[2]

[1]*University of Bahrain*
[2]*University of Bahrain*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025
Accepted: 27 Feb 2025

ABSTRACT

This paper investigates the creation and application of a real-time soft sensor aimed at improving process stability and control by predicting product composition in Crude Distillation Units (CDUs). Creating a machine-learning model capable of continuously analysing and predicting product composition during crude oil distillation is utilized for better operation and control. The research is founded on experimental data obtained from validated dynamic simulations of the CDU process using Aspen-HYSYS software. Time Series Linear Regression (TSLR), Time Series Partial Least Squares (TSPLS), and Time Series Neural Networks (TSNN) are methodologies employed in various soft sensor models developed. Performance metrics, such as root mean square (RMS) and the coefficient of determination (R-squared), facilitate the assessment of these models. The Time-series Neural Network (TSNN) distinguishes itself as the optimal model for estimating distillation endpoints in CDUs, achieving the lowest Root Mean Square (RMS) error of 0.8006 and the highest coefficient of determination (R-squared) due to its superior accuracy and predictive capabilities. The TSNN soft sensor, integrated with the Aspen HYSYS modelling plant, provided real-time estimations of diesel molar flow and composition. The model, integrated with the simulated plant, was trained on real-time data from an Aspen HYSYS simulation of a crude oil distillation unit, enabling continuous live estimations.

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

INTRODUCTION

In the domain of industrial process optimisation, particularly concerning crude oil refining, the pursuit of enhanced efficiency and quality control remains a significant challenge. The initial refining process relies heavily on crude distillation units (CDUs), which segregate crude oil into various fractions based on their boiling points. Optimal performance and adherence to stringent product quality standards rely on precise assessments of product composition during CDU operations.

Contemporary methodologies primarily utilise physical sensors and laboratory analyses to monitor and regulate product compositions in CDUs. While these traditional methods provide valuable analysis, they are occasionally limited by elevated costs, maintenance requirements, and challenges in capturing real-time variations in product compositions. This reliance on offline measurements underscores the necessity for more adaptable and cost-effective technologies capable of functioning optimally within the dynamic environment of a refinery.

The primary objective of this work is to deliver solutions through a real-time soft sensor designed to estimate CDU product compositions. Formulating the mathematical model is a laborious task that could be supplanted by the creation of a soft sensor trained via machine learning. The learning process of the soft sensor depends on the precision of the model created to emulate the industrial plant as a digital twin. These soft sensors would deliver the CDU product composition for oversight and regulation. Historical data from various time zones is utilised to validate the developed model, enhancing the learning process. During training, it is essential to filter information and identify anomalies to ensure the accuracy of a soft sensor. Methodologies will be employed to streamline the soft sensing process.

Soft sensors have been developed specifically for predicting product composition through continuous monitoring of process streams in a refinery crude distillation unit. Partial least squares, artificial neural networks, and linear regression analysis have been utilised in the development of both multiple linear and nonlinear soft sensor models.

Objective of the manuscript

The project's objective is to develop a reliable computational model or algorithm that can accurately predict the distillation endpoint in real-time during the crude oil distillation process in a CDU, utilising a soft sensor that employs advanced mathematical or artificial intelligence techniques to analyse extensive process data and produce precise predictions of the distillation endpoint. When implemented effectively, this technology can substantially enhance the operational efficiency of the crude distillation unit, elevate product quality, and refine process control.

Inspiration

Traditional monitoring methods for ascertaining the distillation endpoint in crude distillation units are both costly and less dependable. Consequently, the refining sector must enhance efficiency, refine process control, and adhere to specifications and regulations. This study will provide a dependable and economical real-time estimation solution via soft sensors, thereby diminishing reliance on conventional monitoring methods and improving the stability and controllability of refining operations. This will enhance productivity in CDUs, as well as product quality and adherence to regulatory standards.

LITERATURE REVIEW

The transformation of crude oil into products like diesel, kerosene, and naphtha constitutes the cornerstone of the energy sector. Crude Distillation Units (CDUs) are integral to this process, requiring precise control systems to maintain efficiency and ensure product quality. Historically, physical sensors have been utilised for measurement and process management within these systems. However, they entail drawbacks, including significant costs for installation and maintenance, as well as delays resulting from reliance on manual sampling and protracted analytics.

The implementation of soft sensors has attracted considerable attention to mitigate these limitations. These data-driven models forecast process variables utilising readily available inputs, providing faster and more cost-effective alternatives to traditional sensors (Park, 2015; Rogina et al., 2011; Ujević Andrijić et al., 2011). Recent advancements in machine learning (ML) and artificial neural networks (ANNs) have markedly improved the capabilities of soft sensors, facilitating real-time monitoring of complex systems (Kubosawa S. & Tsuruoka, 2022; Oster et al., 2023; Shukla et al., 2020; Vedin et al., 2023; Wei et al., 2025; Yoon et al., 2022). This paper analyses enhancements, practical implementations, challenges, and future opportunities for the amalgamation of soft sensors with advanced technologies such as digital twins.

Soft sensors have advanced significantly alongside enhancements in computational methods and algorithms. Preliminary iterations, as highlighted by (Rogina et al., 2011; Ujević Andrijić et al., 2011), were founded on linear and nonlinear modelling techniques. These models depended heavily on data preparation, encompassing outlier detection and dataset normalisation, to improve their predictive effectiveness. With the escalation of system complexity, linear models proved insufficient in accurately depicting the intricate dynamics of CDU operations, thereby requiring the adoption of more advanced nonlinear methodologies.

Kubosawa et al. (2022) illustrated the efficacy of hybrid models that amalgamate dynamic simulations with artificial intelligence methodologies (Kubosawa S. & Tsuruoka, 2022). The amalgamation of methodologies not only enhanced the adaptability of soft sensors to variable operating conditions but also strengthened their robustness. Lüthje et al. (2020) demonstrated the application of hybrid models, integrating data-driven methodologies with mechanical process understanding, for nonlinear predictive control, producing reliable outcomes in various contexts (Lüthje et al., 2020).

Artificial neural networks excel at handling the inherent nonlinearities of CDU systems. Through the analysis of historical data, these networks can predict critical variables, such as feed composition and the quality of end products. GaJang et al. (2010) illustrated the utilisation of feedforward neural networks in delineating complex relationships between input and output variables (GaJang et al., 2010).

Kataria and Singh (2017) made significant advancements by employing recurrent neural networks (RNNs) for temporal data analysis. The ability of RNNs to process sequential data through feedback loops makes them highly appropriate for dynamic processes like crude oil distillation. Their research highlighted the superior effectiveness of RNN-based sensors in environments with variable operating parameters, compared to traditional static models .

Building on this foundation, Park et al. (2015) developed ANN-based soft sensors for real-time feed monitoring, thereby reducing the need for physical analysers (Faruk, 2010; Park, 2015; Perera et al., 2023; Singh & Patel, 2018). The ability to adapt to real-world data and provide actionable insights underscores the importance of artificial neural networks in modern refining operations (Chatterjee & Saraf, 2004; GaJang et al., 2010; Popoola et al., 2013).

In CDU scenarios, where conditions are perpetually evolving, the analysis of time-series data is imperative. Advanced architectures, such as Long Short-Term Memory (LSTM) networks, have exhibited significant effectiveness for specific tasks. Chatterjee and Saraf (2004) examined the application of LSTM models for predicting product compositions in distillation processes, despite the existence of noise or incomplete data. The selective memory retention capabilities of LSTMs ensure accuracy and reliability in these dynamic systems (Chatterjee & Saraf, 2004; Morris et al., 2019).

A principal benefit of soft sensors is their ability to provide real-time insights into process variables. Oster et al. (2023) demonstrated the utilisation of machine learning-based soft sensors in vacuum distillation for the continuous and accurate prediction of product attributes (Oster et al., 2023). This capability enables operators to accelerate informed decision-making, thereby enhancing overall system efficiency and product quality.

Soft sensors are crucial for predictive maintenance and in vacuum distaillation column product estimation (Barbosa, 2014; Niño-Adan et al., 2020; Park, 2015; Ujević Andrijić et al., 2011). These systems reduce unforeseen downtime by examining patterns and detecting early warning signs of equipment degradation. Barbosa (2014) developed a soft sensor to evaluate the quality of hydrocracker products, enhancing process management and providing essential insights into equipment condition. This proactive approach ensures the optimisation of maintenance schedules and minimises disruptions (Barbosa, 2014).

The effectiveness of soft sensors is primarily dependent on the quality of the data used for their training. Inaccurate, deficient, or biassed datasets can significantly affect model accuracy (Shang et al., 2014; Wang & Chen, 2016). To address these challenges, techniques such as noise reduction, feature engineering, and data augmentation should be utilised during the preprocessing phase.

Despite their high accuracy, machine learning models often encounter criticism for lacking interpretability. Operators in critical sectors, such as refining, must depend on the insights provided by these models. Hybrid methodologies that combine domain expertise with data-driven modelling offer a clearer and more understandable framework (Li & Sun, 2023).

Dynamic operational environments require models that can rapidly adjust to changing input data. Kim et al. (2022) examined adaptive learning algorithms capable of self-updating, thereby eliminating the need for frequent manual retraining and ensuring consistent performance across various environments (Kim, 2022).

Digital twins—virtual representations of tangible systems—are revolutionising process optimisation. When integrated with soft sensors, digital twins provide a comprehensive framework for monitoring and predictive analysis.

Li and Sun (2023) demonstrated that the incorporation of these technologies enhances decision-making and process efficiency in refinery operations, enabling the creation of more intelligent and responsive systems (Li & Sun, 2023).

Soft sensors, propelled by advancements in artificial intelligence and machine learning, represent a notable progression in CDU process management (Durrani et al., 2018; Perera et al., 2023; Peterson et al., 2025). They offer real-time monitoring, predictive maintenance, and enhanced operational efficiency. Addressing challenges such as data integrity, transparency, and adaptability will be crucial for maximising their potential. The integration of soft sensors with digital twins highlights their transformative impact, establishing them as a crucial component of innovation in the refining industry.

DESIGN AND EXECUTION

This study employs HYSYS's dynamic simulation model, which is already regulated, to produce data for analysing the dynamic behaviour of a Crude Distillation Unit (CDU) concerning its diesel molar flow. Drawing from literature and Principal Component Analysis (PCA), a group of six pivotal input variables deemed most likely to affect diesel output: column top temperature, column bottom temperature, light gas oil temperature, heavy gas oil temperature, pump-around temperature, and pump-around flow rate.

Every input variable is regulated by a PID controller in the HYSYS model. Disturbances are simulated by executing step changes in the setpoints of these PID controllers and observe the subsequent dynamic response of the CDU's molar flow of diesel. This method produces significant time-series data, illustrating the CDU's performance under different operating conditions influenced by alterations in these essential input variables. Examining this data will be crucial for understanding the intricate relationship between these key factors and the molar flow of diesel, facilitating potential enhancements in future control strategies.

Initially, pertinent data was extracted from the plant database, derived from dynamic simulations conducted in Aspen HYSYS. Six input variables were selected to evaluate their impact on 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) functioned as the output variable. Seven PID controllers were implemented in HYSYS to examine these relationships. Six controllers managed the input variables, while the final controller regulated the diesel molar flow, as illustrated in **Figure 1**. The complete plant simulation is illustrated in **Figure 2**.

The initial PID controller, FIC-102, is designed to regulate the target molar flow rate of the diesel product. The process variable (PV) was initially defined on the "Connections" tab in Fi. This variable denotes the actual molar flow of diesel measured by the plant, which FIC102 will continuously monitor and strive to regulate at the designated setpoint. The output target object (OT) was identified. This denotes the variable that the controller will modify to affect the process variable. In this instance, FIC-102 regulates the actuator position of valve VLV-100 to manage the diesel molar flow.

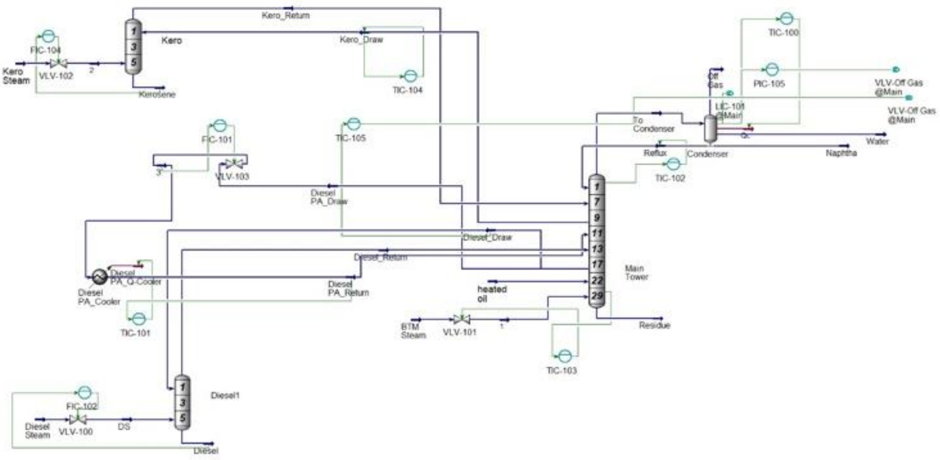
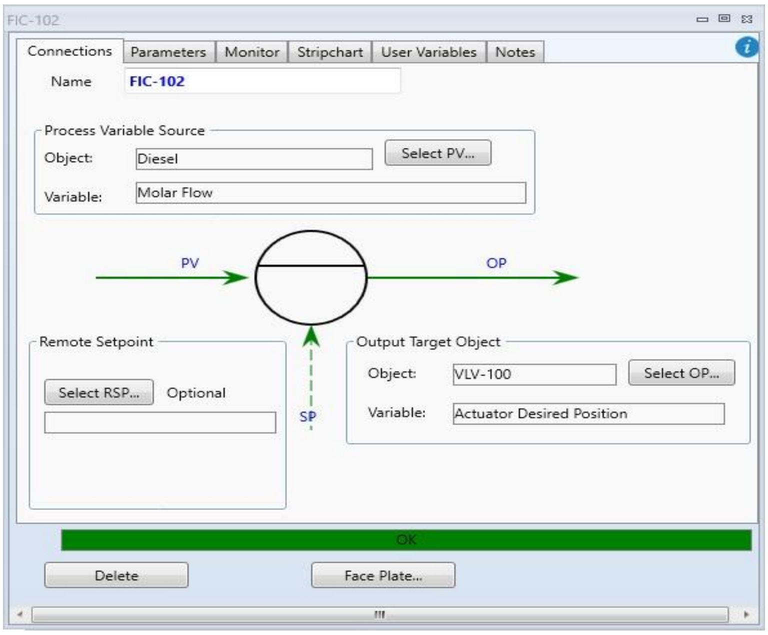


Figure 1. The simulated process flow diagram of the CDU plant



Name: FIC-102

Process Variable Source
Object: Diesel
Variable: Molar Flow

Remote Setpoint
Optional

Output Target Object
Object: VLV-100
Variable: Actuator Desired Position

OK

Figure 2. Connections tab of the FIC-102 controller

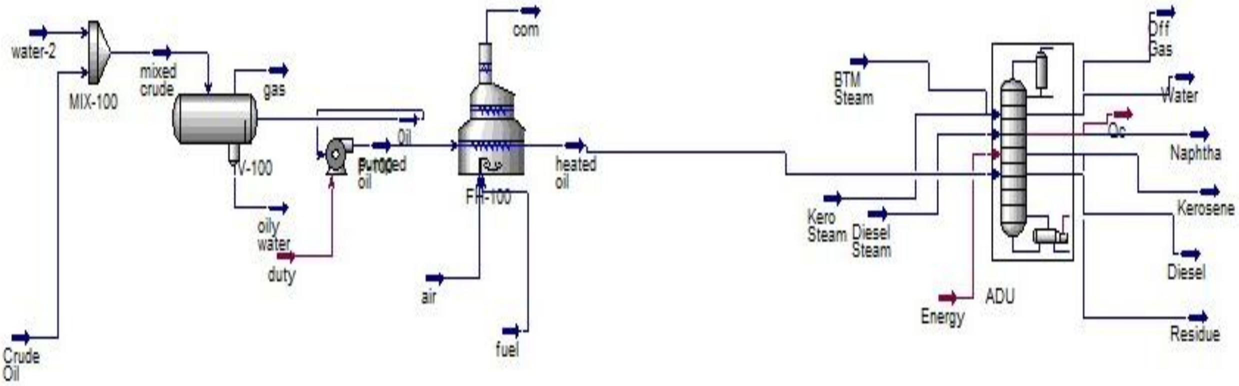


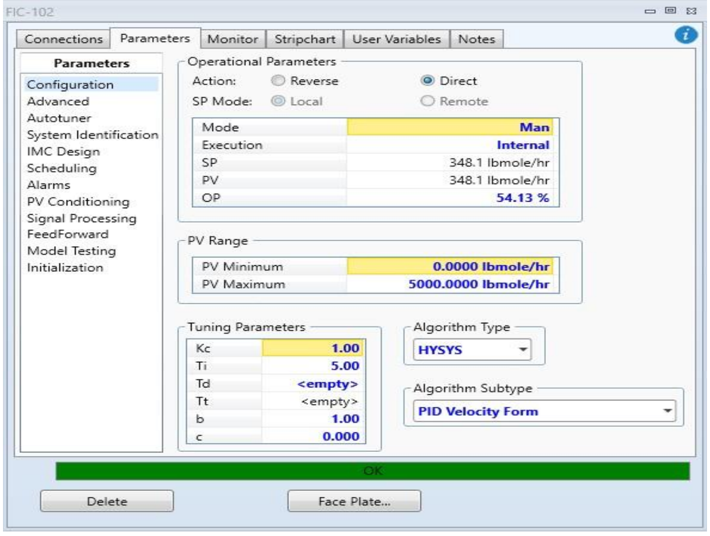
Figure 3. Comprehensive plant simulation

Four operational parameters for FIC-102 were delineated. The action mode was configured to "Direct", signifying a proportional correlation between the controller output and the valve position. Furthermore, the setpoint, representing the target diesel molar flow rate, was established alongside the minimum and maximum permissible values for the process variable. Ultimately, tuning parameters that affect the controller's responsiveness and stability are determined through process knowledge and engineering expertise.

Upon configuration, the controller was engaged by toggling its mode to "Auto" or "Man". The facility is initially operated in manual mode to assess the influence of the input variables on the diesel output. Data collection for analysis was performed through model testing utilising FIC-102. The testing procedure encompassed the delineation of multiple parameters. A step change is implemented to introduce variations in the setpoint. The amplitude of signal variation was ascertained according to the magnitude of this variation. The time interval was designated to generate the desired frequency of data points. The testing duration is established to determine the overall length of the model evaluation.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Operational Parameters

Action: ○ Reverse ● Direct
SP Mode: ● Local ○ Remote

Mode	Man
Execution	Internal
SP	348.1 lbmole/hr
PV	348.1 lbmole/hr
OP	54.13 %

PV Range

PV Minimum	0.0000 lbmole/hr
PV Maximum	5000.0000 lbmole/hr

Tuning Parameters

Kc	1.00
Ti	5.00
Td	<empty>
Tt	<empty>
b	1.00
c	0.000

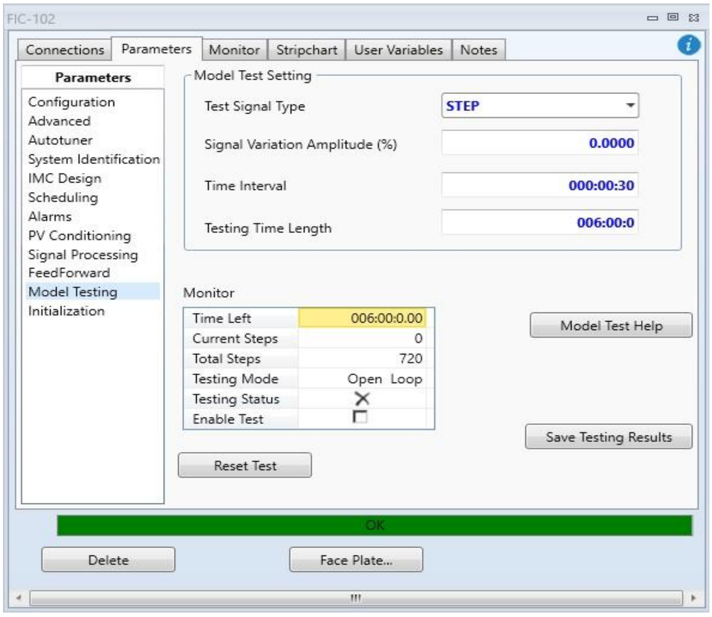
Algorithm Type: HYSYS

Algorithm Subtype: PID Velocity Form

OK

Figure 4. Parameters tab of the FIC-102 controller

Following the establishment of these parameters (**Figure 5**), the simulation was executed for a duration of six hours. The test results were exported to Excel. This data is utilised to construct and train the soft sensor model. The time series plot of the bottom temperature data is illustrated in **Figure 6**. The time series plot of the Diesel Molar Flow is illustrated in **Figure 7**. The initial readings exhibit instability; therefore, they will be excluded from the training process to prevent potential issues with the model if incorporated into the training data. Consequently, this data is eliminated to prevent the construction of an ineffective model.



Model Test Setting

Test Signal Type	STEP
Signal Variation Amplitude (%)	0.0000
Time Interval	000:00:30
Testing Time Length	006:00:0

Monitor

Time Left	006:00:0.00
Current Steps	0
Total Steps	720
Testing Mode	Open Loop
Testing Status	✕
Enable Test	□

OK

Figure 5. Model Testing Tab of the FIC-102 Controller

Data Examination

A correlation matrix is a crucial statistical instrument that elucidates the relationships among variables within a dataset. It comprises a square matrix in which each cell denotes the correlation coefficient between two variables. Correlation coefficients quantify the strength and direction of the linear relationship between variables, with values ranging from -1 to 1. A value of 1 signifies a perfect positive correlation, 0 denotes no correlation, and -1 indicates a perfect negative correlation, while red colours represent a positive correlation and blue signifies a negative correlation, as illustrated in **Figure. 8**.

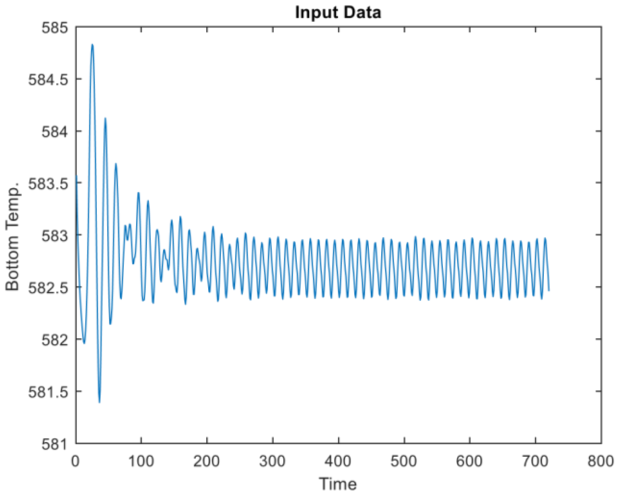


Figure 6. Time series graph of the bottom temperature

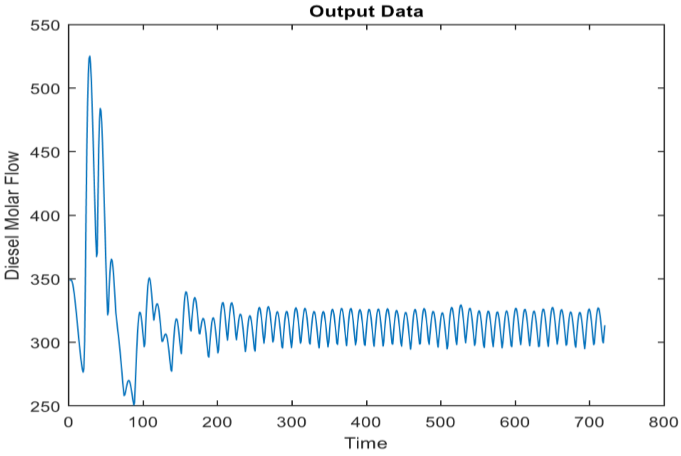
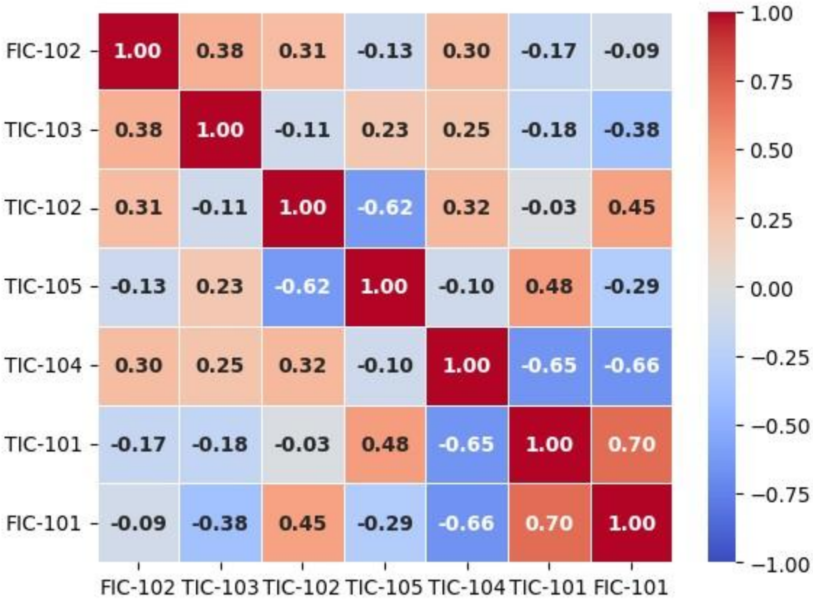


Figure 7. Time Series Graph of the Diesel Molar Flow Rate



	FIC-102	TIC-103	TIC-102	TIC-105	TIC-104	TIC-101	FIC-101
FIC-102	1.00	0.38	0.31	-0.13	0.30	-0.17	-0.09
TIC-103	0.38	1.00	-0.11	0.23	0.25	-0.18	-0.38
TIC-102	0.31	-0.11	1.00	-0.62	0.32	-0.03	0.45
TIC-105	-0.13	0.23	-0.62	1.00	-0.10	0.48	-0.29
TIC-104	0.30	0.25	0.32	-0.10	1.00	-0.65	-0.66
TIC-101	-0.17	-0.18	-0.03	0.48	-0.65	1.00	0.70
FIC-101	-0.09	-0.38	0.45	-0.29	-0.66	0.70	1.00

Figure 8. Heatmap of the correlation matrix

Through the analysis of the correlation matrix in **Figure 8**, we were able to assess the strength of relationships, and comprehend the dependencies among variables. This information is essential for numerous data analysis tasks, including identifying key factors, examining multicollinearity, selecting variables for models, and understanding the dataset's underlying structure. The correlation matrix was used as an effective instrument for summarising and visualising the relationships among variables, serving as a basis for subsequent analysis and interpretation.

The correlation matrix displays the correlation coefficients between the output variable, 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 is 1, indicating a perfect positive correlation.

The correlation coefficients reveal that diesel molar flow (FIC-102) exhibits a weak positive correlation with TIC-102 (0.3063), whereas TIC-103 (0.3784) demonstrates the strongest correlation with the output, suggesting that an increase in the input variable is associated with a slight increase in FIC-102. Conversely, FIC-102 exhibits a weak negative correlation with TIC-105 (0.1284), TIC-101 (-0.1669), and FIC-101 (-0.0914), indicating that as these variables rise, FIC-102 tends to diminish marginally.

Notable correlations exist among the input variables. As observed, TIC-102 and TIC-105 demonstrate a robust negative correlation of -0.6196, whereas TIC-104 and TIC-101 reveal a strong negative correlation of -0.6514. The significant correlations indicate collinearity among these variable pairs. The correlation between bottom temperature and diesel molar flow is illustrated in **Figure 9**.

Cross-correlation analysis is an essential method for examining the relationship between two time series data sets, specifically regarding how fluctuations in one variable may affect another with a potential time lag. The cross-correlation is employed in the simulated case to examine the relationship between the manipulated operating variables, such as column top temperature, and the diesel molar flow during the simulation. This will assist in determining how modifications to these operating conditions may affect diesel production over time. By analysing the temporal displacement at which the correlation between variables reaches its zenith, one can derive insights into the causal relationships inherent in the process.

In the correlation between two time series (x(t) and y(t)), the series was influenced by previous lags of the x-series. The sample cross-correlation function (CCF) is instrumental in determining lags of the x-variable that served as effective predictors of y(t).

The x-variable(s) was designated as the primary variable for the y-variable to forecast future values of y. Consequently, attention is typically directed towards the negative values in the cross-correlation plot, as illustrated in the **Figure 8**.

In **Figure 9** and **Figure 10**, the cross-correlation (CC) plot helped in identifying time lags that exhibit strong correlations with future y values, and these lagged x variables were incorporated into the input data for constructing the machine learning models. Nevertheless, excessive data inclusion will complicate the model; thus, only the three most significant correlations were incorporated.

Model Formulation

The data is first generated in HYSYS and then utilised in the creation of three different types of time series models in MATLAB. The process begins with two linear models, TSMLR and TSPLS, and concludes with the nonlinear TSNN model. The outcomes of each model are evaluated to determine which one has the greatest predictive capability. Consequently, these models are employed to predict the composition of diesel.

Beginning with the TSMLR model, the lagged version of the original data was utilised to enhance the accuracy of our model. Our analysis subsequently incorporated the k-1, k-2, and k-3 lags as predictors, identified via cross-correlation. The dataset was divided into a training set comprising 70% of the original data for coefficient estimation, and a testing set containing the remaining 30% to evaluate the model's performance and ensure its robustness. This allocation allows us to evaluate the effectiveness of our model on data that is both novel and previously unexamined.

The "regress()" function in MATLAB is utilised to construct the predictive model, grounded in advanced regression techniques.

In relation to TSPLS and time series modelling, lagged duplicates of the original data were utilised as predictors. Cross-correlation analysis was conducted to identify potential temporal correlations in the data by selecting delays of k-1, k-2, and k-3. The data was subsequently partitioned into training (70%) and evaluation (30%) sets. The association between the input variables and the objective variable was modelled through partial least squares (PLS) regression utilising the "plsregress" function in MATLAB. Twenty-four latent variables were selected within the PLS framework to effectively encapsulate the underlying structure of the data.

The TSNN architecture was created utilising the "ntstool" software in MATLAB. An iterative method is utilised to ascertain the optimal network architecture. An essential architecture is established with one hidden layer and six neurones, corresponding to the number of input variables. Thus, the predictive efficacy was enhanced by augmenting the quantity of neurones. The final network comprised a solitary hidden layer with ten neurones. This methodology included a three-second time delay as a notable feature. This allowed the network to encode temporal relationships by learning from both the present input and historical values in the time series. The hidden layer utilised the "tanh" activation function to accommodate the modelling of non-linear relationships in the data. The linear activation function was utilised in the output layer to produce continuous output values relevant to regression tasks. The "Levenberg-Marquardt" method, renowned for its effectiveness, was utilised to train the network. To alleviate overfitting, a systematic data partitioning strategy was employed: 70% of the data was designated for training, 10% for validation to evaluate performance during training, and the remaining 20% was set aside for final testing and evaluation.

The aim of fractionator control is to sustain the specified values for the top-draw product endpoints (y1), side-draw product (y2), and bottom reflux temperature (y7). This is accomplished by adjusting 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 subsequently modified through a control loop that employs the hot steam flow rate as a control variable. Furthermore, there are two quantified disturbances in the system: the heat transfer rate of the upper reflux (l1) and the intermediate reflux (l2). These flows extract heat from the system and are subsequently reboiled in different sections of the plant. The statistical parameters of the TSMLR model is showed in **Table 1**. The TSNN architecture was developed as shown in **Figure 11**.

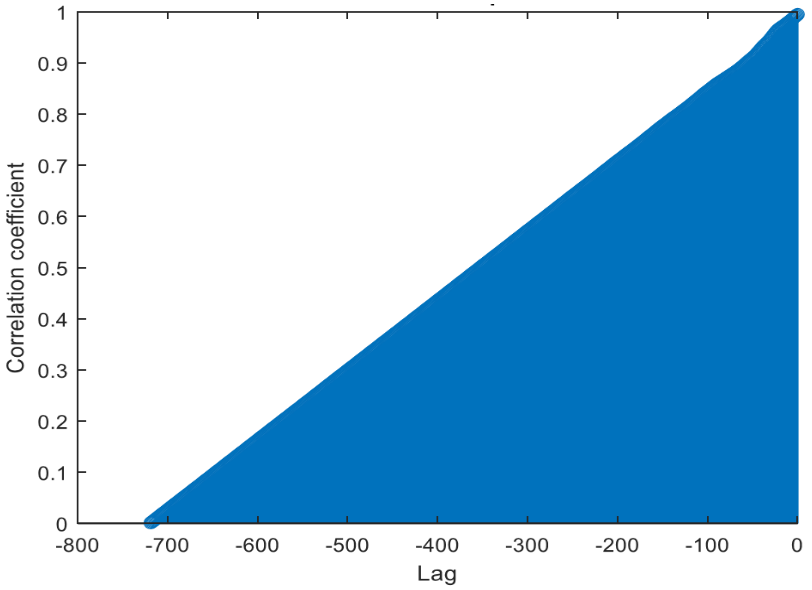


Figure 9. Cross-correlation between bottom temperature TIC-103 and diesel molar flow

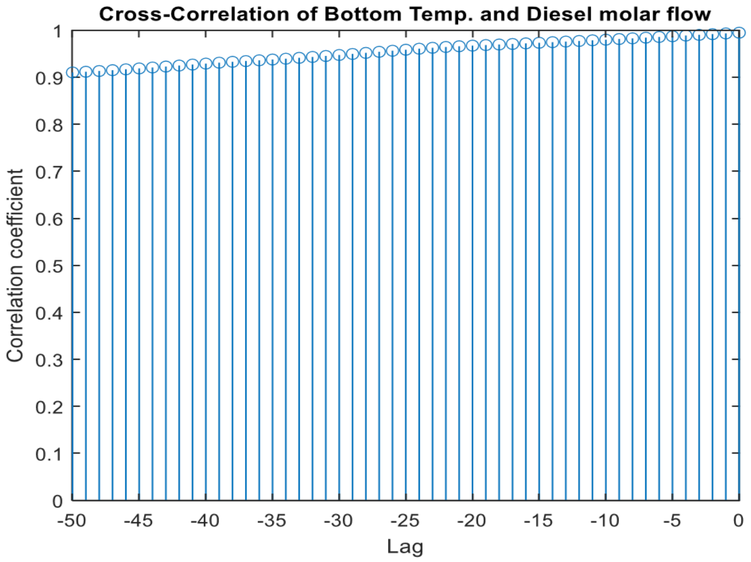


Figure 10. CC plot of bottom temperature versus diesel molar flow

Table 1. Statistical Parameters of the TSMLR Model

Parameters			
R²	***F-test***	***p-value***	***RMSE***
0.7058	6.75	0.0000	11.2509

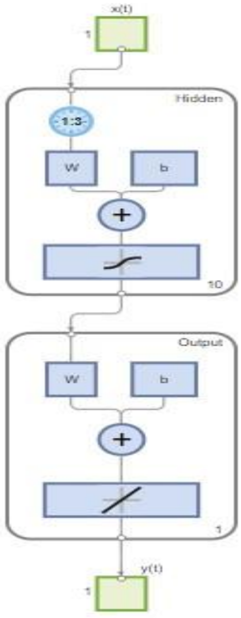


Figure 11. TSNN architecture

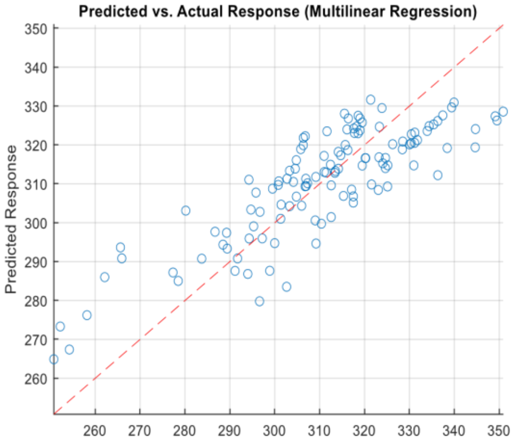


Figure 12. Scatter plot for TSMLR model

FINDINGS AND ANALYSIS

Results of TSMLR Model Development

Table 1 delineates the statistical parameters derived from the analysis of the linear model. The coefficient of determination (R^2) value of 0.7058 signifies a commendable degree of explained variability in the response variable by the predictors. While not exceedingly high, it indicates that the model accounts for a substantial portion of the data's variability. The F-test produced a significant F-value of 6.75 (p-value = 0.0000), signifying the rejection of the null hypothesis of no linear dependence; thus, there is an absence of autocorrelation in the error terms. This validates the existence of a correlation between the predictors and the response variable. The Root Mean Squared Error (RMSE) of 11.2509 indicates the average divergence of the model's predictions from the actual values, implying a satisfactory degree of accuracy. Although the model demonstrates commendable performance, there exists potential for enhancement in accurately capturing variability and minimising prediction errors.

The scatter plot for TSMLR model and the time series plot is illustrated in **Figure 12 and Figure 13**. The model's predictions (orange line) predominantly align with the trend of the actual values (blue line). Nonetheless, there are certain discrepancies between the two lines. This signifies that the model's predictions are not entirely precise, yet they are directionally correct.

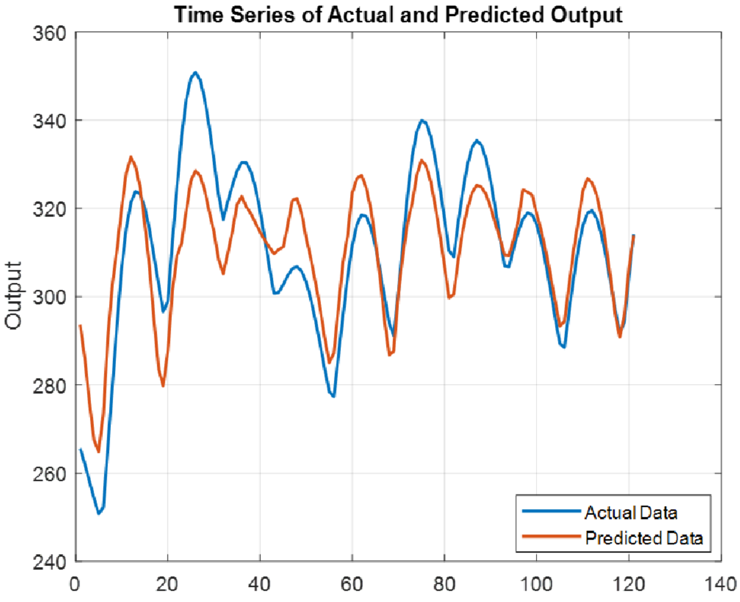


Figure 13. Predictions of the TSMLR model compared to actual data

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Results of TSPLS model development

The Time Series Partial Least Square (TSPLS) model produced highly favourable outcomes, as evidenced by the statistical parameters in **Table 2**. The correlation coefficient (R) between the predicted and actual values was 0.90993, indicating a robust positive linear relationship. This indicates that the TSPLS model successfully identified the fundamental patterns and trends in the time series data. The coefficient of determination (R^2) was determined to be 0.82797, signifying that roughly 82.8% of the variance in the response variable can be elucidated by the predictors incorporated in the model. This illustrates the model's capacity to accurately elucidate and forecast the observed results. The Root Mean Squared Error (RMSE) was calculated as 8.7786, indicating the mean discrepancy between predicted and actual values. A reduced RMSE value indicates an enhanced accuracy, implying that the TSPLS model demonstrated commendable predictive performance. An F-test was performed, yielding an F-value of 7.2193 and a p-value of 6.5281e-14. The substantial F-test demonstrates that the predictors in the model collectively exert a significant influence on forecasting the response variable. The TSPLS model exhibited robust predictive capability, accurately identifying the underlying patterns in the time series data. Referencing **Figure 14** and **Figure 15**, the predicted values closely correspond with the actual values. This alignment signifies that the TSPLS model demonstrates a high degree of accuracy in forecasting and analysing time series data. The closeness of the actual and predicted lines indicates that the model accurately reflects the trends and variations in the data.

Table 2. Statistical Parameters of the TSPLS Model

Parameters			
R²	***F-test***	***p-value***	***RMSE***
0.82797	7.2193	6.5281e^{-14}	8.7786

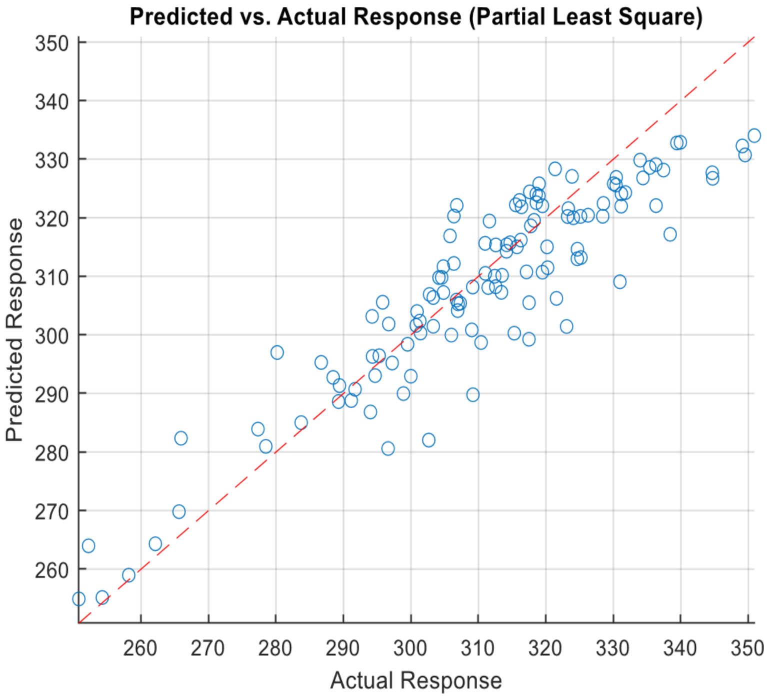


Figure 14. Scatter plot for TSPLS model

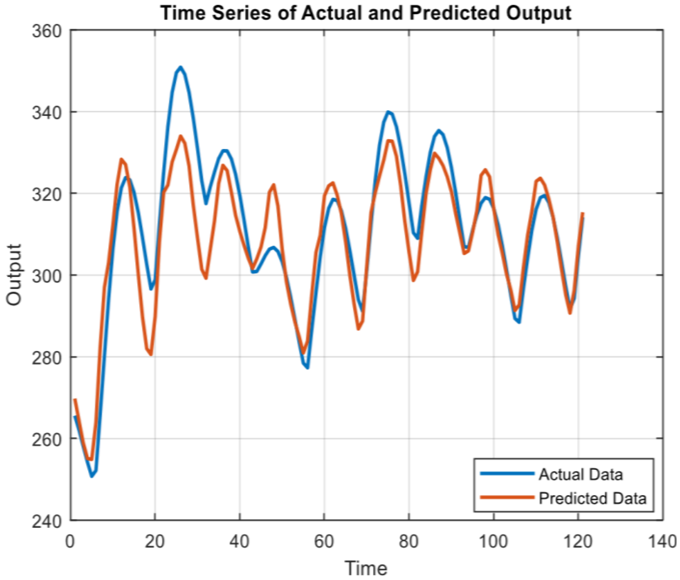


Figure 15. Comparisons of TSPLS model predictions with actual data

Results of TSNN model development

The Time Series Neural Network (TSNN) was the most effective model assessed, attaining the minimal Root Mean Squared Error (RMSE) and maximal R-squared values across all datasets. **Table 3** indicates that it attained the minimal Root Mean Squared Error (RMSE) of 0.1659 during training, signifying exceptional accuracy in estimating product composition in the Crude Distillation Unit (CDU). Moreover, TSNN demonstrated remarkable correlation coefficients, with an R-value of 0.9995 during training and an R^2 value of 0.9997, signifying a robust linear relationship between the predicted and actual values. The results were subsequently corroborated during the validation and testing phases, with TSNN attaining RMSE values of 0.5245 and 0.8008, respectively, alongside elevated R and R^2 values of 0.9982, 0.9991, and 0.9968, 0.9984, respectively.

Additional analysis of the results is presented as illustrated in **Figure 16** and **Figure 17**. **Figure 16** depicts the histogram of target-output errors utilising 20 bins. This study observes that the errors in the plot are confined to a minimal value of 0.007254. All the erroneous training points are in proximity to this point. Thus, the error histogram in this context highlights exceptionally favourable results.

Table 3. Statistical Parameters of the 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 illustrated in **Figure 17** displays iterations indicating the completion of the training process. It delineates the magnitude of the final error and gradient. The training process concluded at the 53rd iteration, which is suboptimal for a trained network. The algorithm selects the 47th iteration due to its lower validation error relative to the training error. This indicates that while the process may yield satisfactory performance on training data through continued iteration, it is not pursued to avert overfitting, which could result in diminished performance. The minimum mean squared error (MSE) is 0.52451 at epoch 74.

This study employs the error autocorrelation function to analyse the temporal interconnection of predictive errors inorder to assess the network's performance. In this instance, Figure 18 displays a solitary nonzero value at zero lag, denoting the mean square error, approximately 0.85, signifying a high degree of accuracy. Aside from the zero-lag correlation, the majority of predictive errors reside within the confidence interval surrounding zero, thereby validating the efficacy of the predictive methodology.

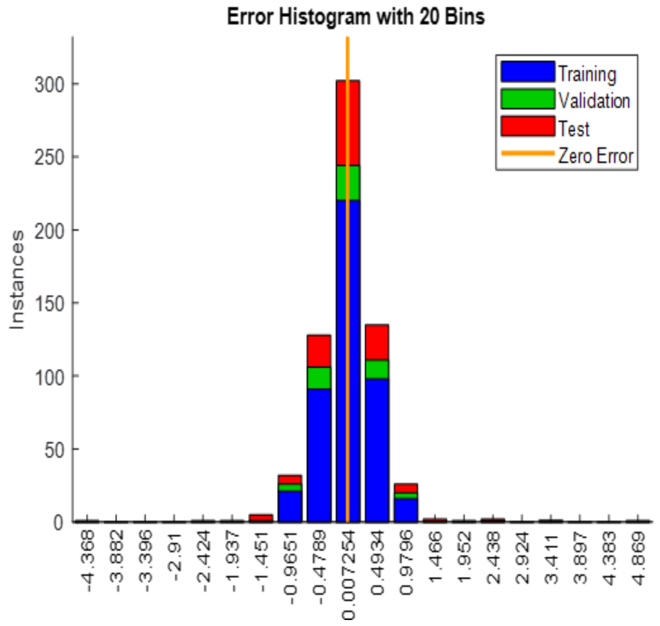


Figure 16. Error histogram comprising 20 bins

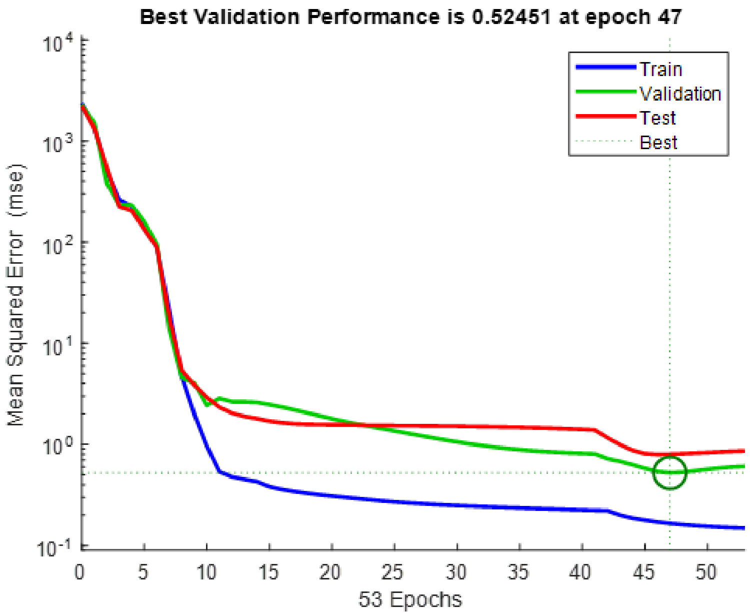


Figure 17. The optimal performance of training, testing, and validation utilising TSNN

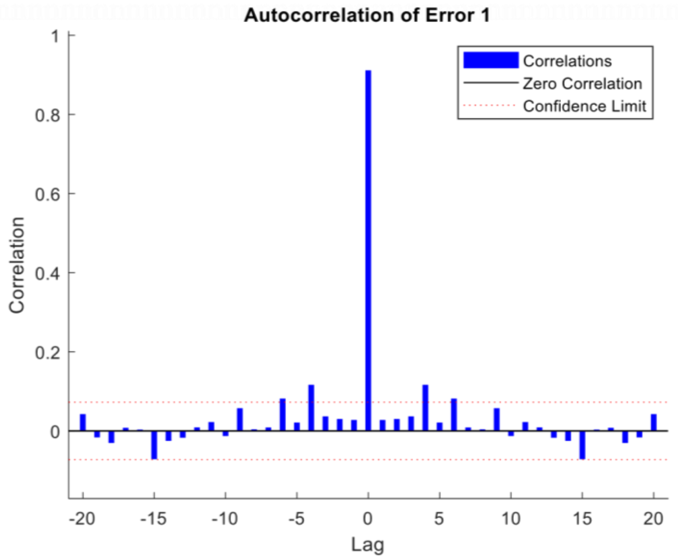


Figure 18. Autocorrelation of residuals

Figure 19 illustrates the model's efficacy in forecasting diesel production. A time series plot illustrates a visual comparison between the model's predictions (outputs) and the actual measured diesel production data (targets). This plot enables the observation of the neural network's response by comparing predicted values with actual values over time. A plot of the prediction errors (the disparity between predicted and actual values) is also included. The model effectively encapsulates the overarching trend of diesel production, illustrating its capacity to depict the desired behaviour. Moreover, the majority of prediction errors reside within the interval of less than -5 to 5, signifying a high degree of accuracy.

The time series neural network consistently demonstrated outstanding performance, attaining remarkable results in further evaluations. **Table 4** displays the results, achieving an RMSE value of 0.1933, signifying a high degree of accuracy in estimating product composition. Furthermore, a substantial R-squared value of 0.9995 is attained, indicating the precision and dependability of TSNN in forecasting product composition. These findings underscore the robustness and efficacy of the TSNN model in practical applications, establishing it as the preeminent performer among the assessed models for arterial intelligence.

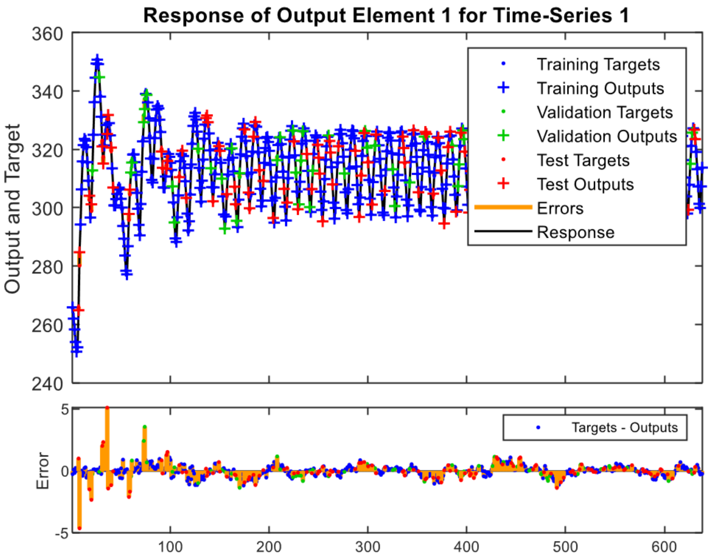


Figure 19. Time series response graph of the target series

Table 4. Supplementary Statistical Parameters for TSNN Model Evaluation

Testing	Parameters		
	RMSE	*R*	*R²*
Additional test	0.1933	0.9990	0.9995

Implementation of real-time soft sensing

The TSNN model was implemented as a real-time soft sensor in Aspen HYSYS utilising three distinct models: TSMLR, TSPLS, and TSNN. The model was seamlessly integrated into the simulation environment, facilitating the continuous estimation of CDU product properties based on real-time measurements of essential input variables. The integration was facilitated by a bespoke MATLAB-HYSYS interface code, serving as a communication conduit between the platforms.

The extensive MATLAB-HYSYS interface code enables real-time data exchange and interaction between the validated soft sensor model in MATLAB and the simulation plant in HYSYS. The integration allows the soft sensor to obtain real-time input data from HYSYS, facilitating precise estimations of the diesel molar flow. **Figure 20** illustrates the predictive accuracy of the real-time soft sensor. where the anticipated values closely align with the actual values, demonstrating the model's remarkable precision in real-time forecasting. The close alignment of the lines indicates that the soft sensor accurately reflects the trends and dynamic variations present in the direct measurement data.

Table 5 presents the RMSE and R values for the three distinct models. The soft sensors exhibited optimal performance with the TSNN model, followed by TSPLS and TSMLR, respectively. Real-time soft sensing was attainable through various methods; however, accuracy varied depending on the developed model. This aligns with the visual results presented in **Figure 20**.

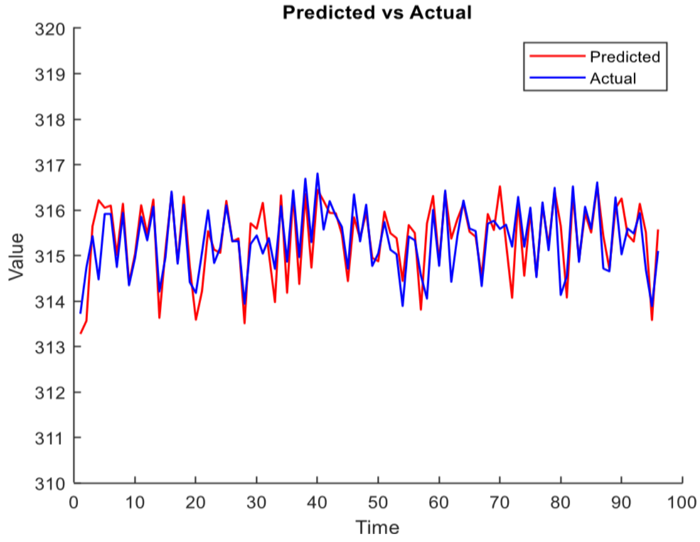


Figure 20. Real-time soft sensor forecasting utilising live measurements of input variables.

Table 5. Performance Evaluation of Soft Sensor Models

Model	Parameters	
	RMSE	*R*
TSMLR	11.2509	0.7058
TSPLS	8.7786	0.82797

Model	Parameters	
	RMSE	***R***
TSNN	0.8006	0.9968

CONFLICT OF INTEREST

There are no conflicts of interest.

CONCLUSIONS

The study aimed to improve product composition monitoring in crude oil distillation units (CDUs) through the development of a real-time soft sensor for estimating distillation endpoints. Traditional methods are costly, inefficient, and lack online metrics. The research successfully created a computational model to predict distillation endpoints in real-time using various machine learning techniques. Dynamic process data was generated from an Aspen HYSYS simulation, subsequently preprocessed and analysed to determine input variables and identify outliers or missing values.

Three soft sensor models were developed using different time-series techniques: Time Series Multi Linear Regression (TSMLR), Time Series Partial Least Squares (TSPLS), and Time Series Neural Network (TSNN). A systematic increase in model complexity was executed, with each subsequent model improving upon its predecessor. This methodology enabled a systematic evaluation and comparison of model effectiveness. The TSMLR model established an initial baseline but exhibited limited accuracy, with a root mean square error (RMSE) of 11.2509. The TSPLS method improved this with an RMSE of 8.7786 by detecting nonlinear interactions through latent variables. The TSNN model produced the most advantageous results, achieving an RMSE of 0.8006 and an R-squared value of 0.9968 during validation with previously untested data. The optimised TSNN soft sensor was then implemented in the Aspen HYSYS environment to demonstrate its ability to provide continuous real-time estimates of the distillation endpoint. The effective deployment of the soft sensor was accomplished by delivering continuous and dependable estimations of product attributes, thereby facilitating monitoring, process regulation, stability improvement, decision-making, and quality assurance.

REFERENCES

[1] Barbosa, J. M. (2014). Development of soft sensors for hydrocracker product quality prediction. Journal of Process Control, 24(2), 150–160.

[2] Chatterjee, S., & Saraf, D. N. (2004). Artificial neural network models for dynamic systems. Chemical Engineering Science, 59(3), 637–649.

[3] Durrani, M. A., Ahmad, I., Kano, M., & Hasebe, S. (2018). An Artificial Intelligence Method for Energy Efficient Operation of Crude Distillation Units under Uncertain Feed Composition. Energies, 11, 12. https://doi.org/10.3390/en11112993

[4] Faruk, D. (2010). A hybrid neural network and ARIMA model for water quality time series prediction. Eng. Appl. of AI, 23, 586–594. https://doi.org/10.1016/j.engappai.2009.09.015

[5] GaJang, H., Lee, K., Kim, M., & Park, S. (2010). ANN-based feed identification in crude distillation. Computers and Chemical Engineering, 34(5), 731–742.

[6] Kim, J. et al. (2022). Adaptive learning algorithms for dynamic process control. Industrial & Engineering Chemistry Research, 61(3), 1234–1242.

[7] Kubosawa S., O. T., & Tsuruoka, Y. (2022). AI-enhanced soft sensors for chemical facilities. AIChE Journal, 68(7), 44–56.

[8] Li, Y., & Sun, X. (2023). Integrating digital twins with soft sensors in refineries. Computers & Chemical Engineering, 172(1), 105–113.

[9] Lüthje, T., Schmidt, M., Weber, P., & Krause, A. (2020). Hybrid models in nonlinear predictive control. Journal of Process Control, 85(6), 128–141.

[10] Morris, D., Davis, P., Sanders, R., & Thompson, K. (2019). Applications of LSTMs in process optimization. Journal of AI Refining, 28(4), 89–102.

[11] Niño-Adan, I., Landa-Torres, I., Manjarres, D., & Portillo, E. (2020). Soft-sensor design for vacuum distillation bottom product penetration classification. Applied Soft Computing, 102, 107072. https://doi.org/10.1016/j.asoc.2020.107072

[12] Oster, J., Braun, M., Meier, L., & Schmidt, P. (2023). ML-based soft sensors in vacuum distillation. Chemical Engineering Research and Design, 179(2), 567–576.

[13] Park, S. J. (2015). Neural network-based software sensors in CDUs. Refinery Systems, 4(5), 77–84.

[14] Perera, Y. S., Ratnaweera, D. A. A. C., Dasanayaka, C. H., & Abeykoon, C. (2023). The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review. Engineering Applications of Artificial Intelligence, 121, 105988. https://doi.org/https://doi.org/10.1016/j.engappai.2023.105988

[15] Peterson, L., Gosea, I. V., Benner, P., & Sundmacher, K. (2025). Digital twins in process engineering: An overview on computational and numerical methods. Computers & Chemical Engineering, 193, 108917. https://doi.org/https://doi.org/10.1016/j.compchemeng.2024.108917

[16] Popoola, L. T., Babagana, G., & Susu, A. A. (2013). Expert system design and control of crude oil distillation column of a Nigerian refinery using artificial neural network model.

[17] Rogina, A., Šiško, I., Mohler, I., Andrijić, Ž., & Bolf, N. (2011). Soft sensor for continuous product quality estimation (in crude distillation unit). Chemical Engineering Research & Design - CHEM ENG RES DES, 89, 2070–2077. https://doi.org/10.1016/j.cherd.2011.01.003

[18] Shang, C., Yang, F., Huang, D., & Lyu, W. (2014). Data-driven soft sensor development based on deep learning technique. Journal of Process Control, 24, 223–233. https://doi.org/10.1016/j.jprocont.2014.01.012

[19] Shukla, R., Verma, P., Agarwal, M., & Gupta, N. (2020). Dimensionality reduction in ML-based soft sensors. AI in Chemical Engineering, 62(3), 89–102.

[20] Singh, R., & Patel, D. (2018). Time-series ML in CDU modeling. Chemical Engineering Review, 45(6), 221–234.

[21] Ujević Andrijić, A. Ž., Mohler, I., & Bolf, N. (2011). Development of soft sensors for CDU control. Refining Technology, 3(4), 231–239.

[22] Vedin, A. E., Odemis, S., Sener, A., Kayar, G., & Aliyev, M. (2023). Unsupervised anomaly detection model for diesel off-spec color change triggered by flooding. In A. C. Kokossis, M. C. Georgiadis, & E. Pistikopoulos (Eds.), Computer Aided Chemical Engineering (Vol. 52, pp. 1873–1878). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-443-15274-0.50297-3

[23] Wang, L., & Chen, Y. (2016). The role of data preprocessing in ML. AIChE Journal, 62(7), 177–186.

[24] Wei, B., Zhang, B., Che, L., Lin, H., & Zhou, H. (2025). Novel integrated EMPC framework for fluid catalytic cracking unit under feedstock uncertainty. Chemical Engineering Research and Design, 222, 108–120. https://doi.org/https://doi.org/10.1016/j.cherd.2025.09.001

[25] Yoon, Y. S., Jeong, W., Kim, J., Seok, M., Park, J., Bae, J., Lee, K., & Lee, J. H. (2022). Development of inferential sensor and real-time optimizer for a vacuum distillation unit by recurrent neural network modeling of time series data. Computers & Chemical Engineering, 168, 108039. https://doi.org/https://doi.org/10.1016/j.compchemeng.2022.108039