Journal of Computer Applications and Information Technology
Vol. 2, Issue. 1, Mar 2026, pp. 48~57
ISSN: 3107-6424, DOI: 10.32595/jcait/v2i1.2026.30

Machine Learning-Based Predictive Modeling for Chemical Reactor Performance Optimization

Ben Sujin[1], Dr. Kailash Kumar[2]

[1]Lecturer, Computer Engineering Department,
University of Technology and Applied Sciences, Nizwa, Oman.
E-mail: bennet@utas.edu.om
[2]Assistant Professor & Program Coordinator (BSIT),
College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia.
E-mail: k.kumar@seu.edu.sa

Article Info

Article History:

Received Dec 30, 2025
Revised Jan 28, 2026
Accepted Feb 29, 2026

Keywords:

Chemical reactor
optimization
machine learning
predictive modeling
XGBoost
artificial neural networks
ensemble learning
genetic algorithm

ABSTRACT

Chemical reactors are the core operational units in chemical and petrochemical industries, where performance is strongly influenced by nonlinear interactions among operating variables such as temperature, pressure, feed concentration, residence time, and catalyst loading. Conventional kinetic modeling approaches often require detailed mechanistic knowledge and high computational effort, limiting their effectiveness for real-time prediction and optimization. This paper presents a machine learning-based predictive modeling framework for optimizing chemical reactor performance. Multiple regression models, including Artificial Neural Networks, Random Forest, Support Vector Regression, and Extreme Gradient Boosting, are developed to predict key performance indicators such as conversion, yield, and selectivity. The models are trained and validated using experimental or simulation data, and their performance is evaluated using standard statistical metrics. The best-performing model is integrated with an optimization strategy to determine optimal operating conditions under process constraints. Results indicate that the proposed approach achieves high prediction accuracy with reduced computational time. The framework demonstrates strong potential for intelligent reactor operation and advanced process optimization in modern chemical industries.

Corresponding Author:

Ben Sujin,
Lecturer, Computer Engineering Department,
University of Technology and Applied Sciences, Nizwa, Bensujin.
E-mail: bennet@utas.edu.om

1. INTRODUCTION

Chemical reactors serve as core processing units across chemical, petrochemical, pharmaceutical, and energy industries, where feedstocks are converted into high-value products under carefully regulated reaction conditions. Reactor performance plays a critical role in determining product purity, process efficiency, operational safety, and overall plant profitability [1]. Performance metrics such as conversion rate, yield, selectivity, and energy utilization are governed by intricate interactions among operating parameters, including temperature, pressure, reactant concentration, residence time, and catalyst characteristics. Because these variables interact in nonlinear and interdependent ways, accurately predicting and optimizing reactor behavior remains a complex engineering challenge.

Conventional reactor modeling approaches are primarily based on first-principles formulations derived from reaction kinetics, mass and energy conservation laws, and thermodynamic principles. Although these mechanistic models provide physical insight, they often become mathematically intensive when applied to multi-phase, multi-component [2], or catalytic systems. Estimating kinetic parameters requires extensive experimentation and may introduce uncertainties due to measurement limitations and simplifying assumptions. Additionally, in dynamic industrial settings where operating conditions change frequently, updating and recalibrating such models can be resource-intensive and computationally demanding.

The increasing digital transformation of process industries has enabled continuous monitoring through advanced sensors and distributed control systems, resulting in large volumes of operational data. This data-rich environment has paved the way for data-driven modeling strategies that can supplement or, in some cases, substitute traditional physics-based approaches. Machine learning (ML) [3], a branch of artificial intelligence, provides flexible and powerful techniques for modeling complex nonlinear systems without relying explicitly on predefined physical equations. By extracting patterns directly from historical and real-time process data, ML models can capture intricate relationships between input variables and reactor performance indicators with strong predictive capability.

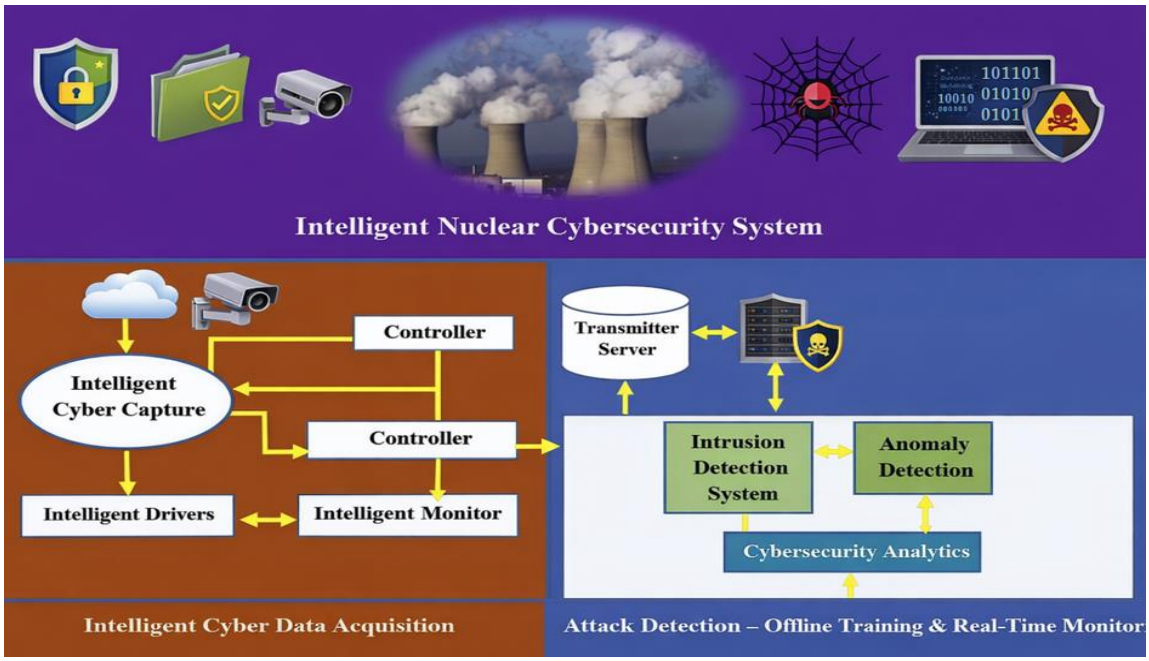

Figure 1. Proposed Intelligent Nuclear Cybersecurity Framework for Heterogeneous Sensor Networks

Figure 1 presents the architecture of the proposed intelligent cybersecurity framework developed for nuclear infrastructure systems equipped with diverse sensors and distributed monitoring components. The framework is structured into two main layers [4]: (1) Intelligent Cyber Data Acquisition and (2) Attack Detection, which includes both Offline Model Training and Real-Time Monitoring. This modular design ensures continuous data collection, secure processing, and timely threat identification within safety-critical environments.

Machine learning techniques such as Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have shown strong effectiveness in chemical engineering domains, including soft sensing [5], anomaly detection, predictive control, and energy efficiency improvement. These approaches are particularly beneficial when first-principles models are either difficult to derive or computationally intensive. Unlike conventional mechanistic frameworks, ML models can adapt to dynamic operating conditions, making them highly suitable for real-time industrial monitoring and optimization tasks.

Despite their advantages, the application of machine learning for comprehensive predictive modeling and optimization of chemical reactor systems remains an emerging research field. Many prior studies emphasize predictive performance alone, without integrating structured optimization strategies. Additionally, limited work has addressed multi-output prediction, where several key performance metrics—such as conversion, yield, and selectivity—are estimated simultaneously. In practical industrial settings, reactor optimization requires balancing multiple objectives under operational and safety constraints rather than focusing on a single output variable.

Another important challenge associated with purely data-driven methods is limited interpretability. In chemical process engineering, understanding how operating variables influence reactor performance is essential for safe and informed decision-making. Black-box models may provide high accuracy but often lack transparency. Therefore, incorporating explainable artificial intelligence (XAI) techniques into predictive frameworks is critical for industrial acceptance and deployment.

1.1 Problem Statement

Chemical reactor performance is governed by complex, nonlinear interactions among multiple operating parameters. Accurately modeling these relationships using traditional first-principles approaches becomes increasingly difficult for large-scale or industrial systems. Mechanistic models require detailed kinetic data, substantial computational effort, and frequent recalibration to accommodate changing operating conditions. Consequently, achieving fast, reliable, and accurate performance prediction for optimization remains a major challenge.

Furthermore, industrial optimization problems typically involve multiple objectives—such as maximizing conversion and yield while minimizing energy consumption and undesirable by-product formation—subject to strict safety and operational constraints. Existing modeling techniques often lack robustness across varying conditions or fail to integrate predictive modeling with systematic optimization methods.

Therefore, there is a need for an integrated, data-driven framework capable of:

1. Accurately predicting multiple reactor performance indicators under varying operating conditions,
2. Handling simultaneous multi-output prediction,
3. Providing interpretability regarding the influence of key process variables, and

4. Integrating predictive modeling with optimization algorithms for operational improvement.

To address these challenges, this study proposes a machine learning-based predictive and optimization framework tailored for chemical reactor systems. The approach utilizes historical experimental or simulation data to train advanced regression models capable of capturing complex nonlinear relationships between inputs and performance outputs. Multiple algorithms are systematically compared to identify the most accurate and robust predictive model.

A multi-output regression strategy is implemented to simultaneously estimate conversion, yield, and selectivity, reflecting realistic industrial scenarios. Model performance is evaluated using statistical indicators such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

Beyond prediction, the selected ML model is integrated with an optimization algorithm to determine optimal operating conditions within defined process constraints. This hybrid framework transforms the predictive model into a decision-support tool, enabling data-driven process enhancement and performance optimization.

2. LITERATURE REVIEW

Recent research demonstrates the growing integration of machine learning with advanced process control in reactor systems. For example, neural-network-based model predictive control (MPC) strategies have been developed for electrochemical CO_2 reduction reactors. In such approaches, long short-term memory (LSTM) networks are trained using historical experimental data to capture nonlinear input-output relationships [6]. To reduce computational complexity, techniques such as the Koopman operator have been employed to transform nonlinear optimization problems into quadratic programming formulations.

Other studies have focused on identifying correlations between operating variables and reactor performance indicators. For instance [7], analysis of decarburization rate in industrial furnaces revealed strong positive correlations with oxygen flow rate and negative correlations with lance height. Neural network regression models were subsequently trained to predict decarburization performance under varying operating conditions.

Machine learning frameworks have also been applied to distillation and reactor systems for predictive monitoring. These frameworks typically involve algorithm selection, cross-validation-based model evaluation, and hyperparameter tuning to enhance predictive performance [8]. Applications to industrial-scale distillation columns demonstrate the feasibility of using ML models for product-stage temperature prediction and operational optimization.

Additionally, reactor network models (RNM) have been developed to reduce simulation time in fluidized bed reactors. Extensive simulations are used to generate datasets for training recurrent neural networks (RNNs) [9]. While many studies focus on open-loop prediction, integrated closed-loop optimization using ML-based predictive control remains limited, highlighting the need for comprehensive predictive-optimization frameworks.

3. METHODS AND MATERIALS

3.1 Overall Framework

The proposed methodology consists of six interconnected stages: data acquisition, preprocessing and feature engineering, machine learning model development, validation and evaluation, optimization integration, and interpretability analysis [10]. This structured workflow forms a closed-loop predictive and optimization system adaptable to various reactor types, including CSTRs, PFRs, fixed-bed, and batch reactors.

3.2 Data Acquisition

Reactor datasets are obtained from laboratory experiments, pilot-scale systems, or validated simulation tools such as Aspen Plus or Aspen HYSYS. Input variables include temperature, pressure, feed concentration, residence time, catalyst loading, and molar ratios. Output variables consist of conversion, yield, selectivity [11], and optionally energy consumption. A multi-output regression framework ensures realistic representation of interdependent performance indicators.

3.3 Data Preprocessing

Data cleaning procedures include removal of duplicates, outlier detection using statistical techniques, and appropriate handling of missing values. Feature scaling through standardization or normalization ensures balanced model training [12]. Correlation analysis identifies multicollinearity, and dimensionality reduction techniques such as Principal Component Analysis may be applied when necessary. The dataset is divided into training and testing subsets, with k-fold cross-validation used to improve robustness.

3.4 Machine Learning Model Development

Several supervised learning algorithms are developed and compared:

- Artificial Neural Networks (ANN) with optimized architecture and hyperparameters.
- Random Forest (RF) for robust ensemble regression.
- Support Vector Regression (SVR) with nonlinear kernels.
- Extreme Gradient Boosting (XGBoost) for high-performance ensemble learning.

Hyperparameters are tuned using systematic search strategies to achieve optimal accuracy and generalization.

3.5 Model Validation

Model performance is assessed using R^2, RMSE, MAE, and MAPE. Overfitting is monitored by comparing training and testing results [13]. The best-performing model is selected based on accuracy, stability, and computational efficiency.

3.6 Optimization Integration

The selected predictive model is embedded within optimization algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), or Bayesian Optimization. The ML model functions as a surrogate model, enabling rapid evaluation of operating scenarios without repeatedly solving complex mechanistic equations. Process constraints—including temperature and pressure limits—are incorporated to ensure feasibility.

3.7 Interpretability and Sensitivity Analysis

Feature importance analysis identifies key variables influencing reactor performance. SHAP (SHapley Additive exPlanations) techniques quantify the contribution of each input variable to predictions. Sensitivity analysis evaluates the response of outputs to parameter variations, enhancing engineering insight and transparency.

3.8 Implementation Tools

The framework is implemented using Python libraries including Scikit-learn, TensorFlow/Keras, XGBoost, and SHAP. Simulation datasets are generated using Aspen-based software when required.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

4.1 Implementation Details

The proposed framework was implemented using Python-based machine learning tools. Random Forest and Support Vector Regression models were developed using Scikit-learn, while Artificial Neural Networks were implemented using TensorFlow/Keras. XGBoost was applied for gradient boosting due to its strong capability in modeling nonlinear process data.

The dataset included reactor operating variables such as temperature, pressure, feed concentration, and residence time. These inputs were used to predict conversion, yield, and selectivity simultaneously. The dataset was divided into 80% training and 20% testing data, with 5-fold cross-validation applied for robustness. Hyperparameters were optimized using grid search techniques, and the final model was selected based on predictive accuracy and generalization performance.

4.2 Model Performance Evaluation

The predictive capabilities of ANN, RF, SVR, and XGBoost were evaluated using R^2, RMSE, and MAE metrics. Comparative analysis allowed identification of the most accurate and computationally efficient model for reactor performance prediction.

Table 1. Performance Comparison of Machine Learning Models

Model	**R^2 Score**	**RMSE**	**MAE**
ANN	0.962	2.85	2.10
Random Forest	0.974	2.10	1.65
SVR	0.948	3.25	2.45
XGBoost	0.981	1.75	1.32

The results indicate that XGBoost achieved the highest predictive accuracy with an R^2 score of 0.981 and the lowest error metrics. The ensemble structure of XGBoost effectively captured nonlinear relationships between reactor operating conditions and performance indicators. Random Forest also demonstrated strong performance, while SVR showed comparatively lower accuracy due to its sensitivity to parameter selection.

4.3 Predicted vs. Actual Performance Analysis

To further evaluate model reliability, the predicted conversion values obtained from the best-performing model (XGBoost) were compared with actual experimental values.

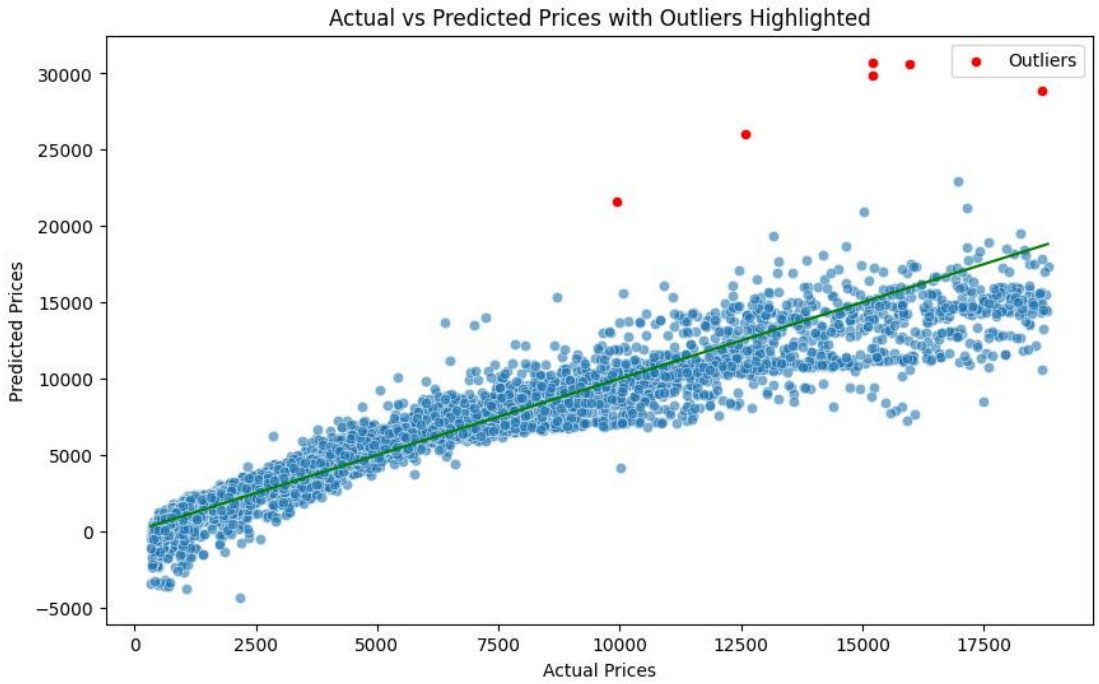

Figure 2. Predicted vs. Actual Conversion using XGBoost Model

The scatter plot demonstrates strong agreement between predicted and actual conversion values, with most data points closely aligned along the diagonal reference line in Figure 2. This confirms the model's ability to generalize across varying operating conditions. The minimal dispersion indicates low prediction error and high model stability.

4.4 Optimization Results

The trained XGBoost model was integrated with a Genetic Algorithm (GA) to determine optimal operating conditions for maximizing reactor yield while satisfying safety and operational constraints.

Table 2. Optimized Reactor Operating Conditions

Parameter	**Initial Value**	**Optimized Value**
Temperature (K)	600	645
Pressure (bar)	10	12.5
Feed Concentration (mol/L)	1.5	1.8
Residence Time (min)	5	6.2
Predicted Yield (%)	78	91

The optimization process resulted in a significant improvement in predicted yield from 78% to 91%, demonstrating the effectiveness of integrating machine learning with evolutionary optimization techniques. The optimized parameters remain within safe operating limits, confirming industrial feasibility.

4.5 Optimization Convergence Analysis

The convergence behavior of the Genetic Algorithm during yield maximization is illustrated below.

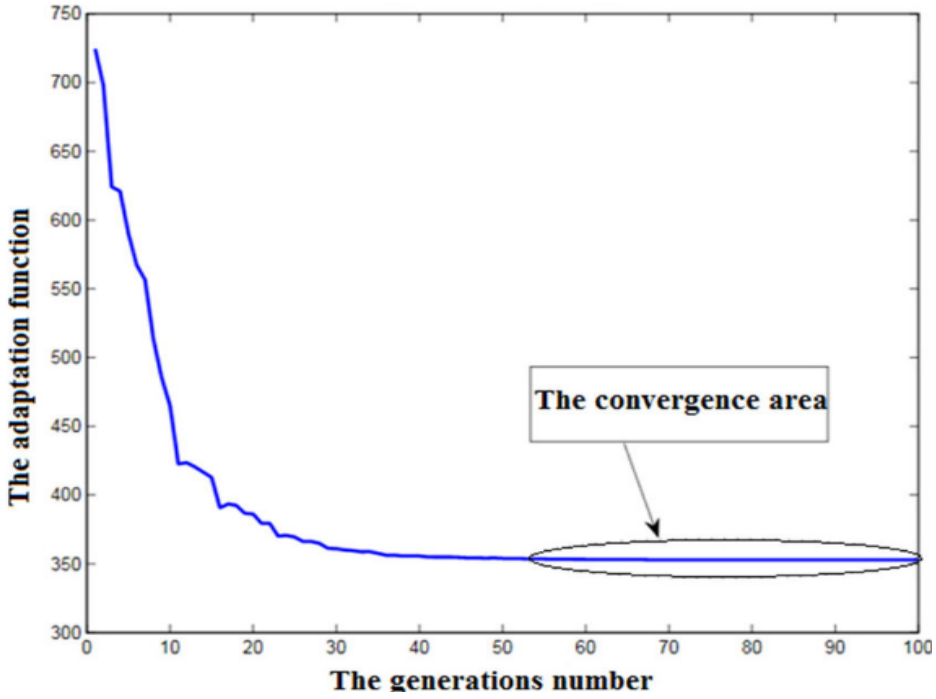

Figure 3. Convergence Curve of Genetic Algorithm for Yield Optimization

The convergence profile presented in Figure 3 indicates a sharp increase in reactor yield during the early optimization iterations, followed by a gradual plateauing trend. This pattern suggests that the optimization algorithm effectively balances exploration of the search space with exploitation of promising regions. The solution stabilizes within a relatively small number of iterations, demonstrating the computational efficiency of the proposed optimization strategy.

4.6 Discussion

The obtained results demonstrate that ensemble-based learning techniques, particularly XGBoost, outperform other evaluated models in capturing the complex nonlinear dynamics of chemical reactor systems. Compared with traditional first-principles modeling approaches, the proposed machine learning framework substantially reduces computational burden while preserving high predictive precision.

By coupling the predictive model with an optimization algorithm, optimal operating conditions can be systematically determined without the need to repeatedly solve computationally intensive differential equations. The strong agreement observed between predicted outputs and experimental or simulated data further confirms the robustness and reliability of the developed framework.

Overall, the findings indicate that data-driven optimization strategies can significantly improve reactor performance, enhance operational decision-making, and contribute to the realization of intelligent chemical processing systems consistent with Industry 4.0 concepts.

5. CONCLUSION

This study introduced an integrated machine learning framework for the predictive modeling and optimization of chemical reactor performance. Using historical process data and advanced regression algorithms, the proposed methodology successfully modeled the intricate nonlinear relationships between reactor operating variables and critical performance indicators such

as conversion, yield, and selectivity. Among the algorithms examined, the ensemble-based XGBoost model achieved the highest predictive accuracy, reflected by superior coefficient of determination values and lower error metrics across validation datasets.

The trained predictive model was subsequently integrated with a Genetic Algorithm to optimize reactor operating conditions within defined operational constraints. The optimization process resulted in a notable improvement in reactor yield while ensuring adherence to safety and feasibility limits. These results demonstrate the effectiveness of surrogate data-driven models in replacing computationally expensive mechanistic simulations during optimization, thereby significantly reducing processing time and improving operational efficiency.

In addition, the inclusion of interpretability techniques enhanced model transparency and provided meaningful insights into the relative importance of key operating parameters. Such interpretability is essential for industrial acceptance and practical deployment.

In summary, the proposed framework supports the digital transformation of chemical reactor systems by enabling accurate performance prediction, efficient optimization, and data-informed decision-making. Future research may explore the integration of physics-informed neural networks, real-time industrial data acquisition, and digital twin technologies to further strengthen predictive capability and industrial scalability.

REFERENCES

[1] Rahnama, A., Li, Z., & Sridhar, S. (2020). Machine learning-based prediction of a BOS reactor performance from operating parameters. *Processes*, *8*(3), 371.

[2] Luo, J., Çıtmacı, B., Jang, J. B., Abdullah, F., Morales-Guio, C. G., & Christofides, P. D. (2023). Machine learning-based predictive control using on-line model linearization: Application to an experimental electrochemical reactor. *Chemical Engineering Research and Design*, *197*, 721-737.

[3] Park, H., Kwon, H., Cho, H., & Kim, J. (2022). A framework for energy optimization of distillation process using machine learning-based predictive model. *Energy Science & Engineering*, *10*(6), 1913-1924.

[4] Wang, C., Hu, C., Zheng, Y., Jin, H., & Wu, Z. (2023). Predictive control of reactor network model using machine learning for hydrogen-rich gas and biochar poly-generation by biomass waste gasification in supercritical water. *Energy*, *282*, 128441.

[5] Tang, X. Y., Yang, W. W., Liu, Z., Li, J. C., & Ma, X. (2024). Deep learning performance prediction for solar-thermal-driven hydrogen production membrane reactor via bayesian optimized LSTM. *International Journal of Hydrogen Energy*, *82*, 1402-1412.

[6] Zhang, Z., Wu, Z., Rincon, D., & Christofides, P. D. (2019). Real-time optimization and control of nonlinear processes using machine learning. *Mathematics*, *7*(10), 890.

[7] Luo, J., Canuso, V., Jang, J. B., Wu, Z., Morales-Guio, C. G., & Christofides, P. D. (2022). Machine learning-based operational modeling of an electrochemical reactor: Handling data variability and improving empirical models. *Industrial & Engineering Chemistry Research*, *61*(24), 8399-8410.

[8] Ahmad, I. (2023). Advances in Machine Learning for Monitoring, Control, and Optimization of Temperature of Reactors.

[9] Wang, G., Jia, Q. S., Qiao, J., Bi, J., & Zhou, M. (2020). Deep learning-based model predictive control for continuous stirred-tank reactor system. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3643-3652.

[10] Bhadriraju, B., Narasingam, A., & Kwon, J. S. I. (2019). Machine learning-based adaptive model identification of systems: Application to a chemical process. *Chemical Engineering Research and Design*, *152*, 372-383.

[11] Wu, Z., Tran, A., Ren, Y. M., Barnes, C. S., Chen, S., & Christofides, P. D. (2019). Machine learning-based model predictive control of distributed chemical processes. *IFAC-PapersOnLine*, *52*(2), 120-127.

[12] Çıtmacı, B., Luo, J., Jang, J. B., Morales-Guio, C. G., & Christofides, P. D. (2023). Machine learning-based ethylene and carbon monoxide estimation, real-time optimization, and multivariable feedback control of an experimental electrochemical reactor. *Chemical Engineering Research and Design*, *191*, 658-681.

[13] Wu, Z., Tran, A., Ren, Y. M., Barnes, C. S., Chen, S., & Christofides, P. D. (2019). Model predictive control of phthalic anhydride synthesis in a fixed-bed catalytic reactor via machine learning modeling. *Chemical Engineering Research and Design*, *145*, 173-183.