ISEIS

Journal of Environmental Informatics Letters 12(2) 79-88 (2024)

Journal of Environmental Informatics Letters

www.iseis.org/jeil

Machine Learning Framework for the Methyl Chloride Production Process

R. Gollangi[1], K. NagamalleswaraRao[2], S. R. Dhanushkodi[2 *], and N. Mahinpey[3]

[1] *School of Mechanical Engineering, Vellore Institute of Technology, Tamil Nadu, Vellore 632014, India*
[2] *Dhanushkodi Research group, School of Chemical Engineering, Vellore Institute of Technology, Tamil Nadu, Vellore 632014, India*
[3] *Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta T2N 1N4, Canada*

Received 20 February 2024; revised 14 June 2024; accepted 02 September 2024; published online 22 October 202

ABSTRACT. We report a framework for calculating the exergy and energy losses in the hydrochlorination of the methanol process for a methyl chloride production unit. Machine learning-based predictive maintenance models are identified to assess plant toxicity. The proposed novel framework integrates Hyprotech Systems (HYSYS) and machine learning models. It optimizes the operating conditions of the chloromethane plant. Both supervised and unsupervised machine learning models such as Bayesian Ridge regression (BRR), Nearest Neighbors regression (KNR), and Stochastic gradient descent regression (SGD) are used to forecast energy and exergy destruction. Among these models, the BRR exhibits exceptional accuracy in predicting thermodynamic losses, notably achieving an R^2 value of 0.998. Thereby, an integrated framework could serve as a valuable diagnostic tool for assessing real-time plant data to enhance plant operational efficiency.

Keywords: energy loss, exergy destruction, machine learning framework, methyl chloride production process, Bayesian Ridge regression

1. Introduction

Methyl chloride is a chemical utilized in the production of silicones, agricultural chemicals, amines, freon, and butyl rubber. It serves as both a solvent and functional fluid in manufacturing industries but is also noted for its carcinogenic properties, as documented in the Toxics Release Inventory. Additionally, methyl chloride is classified as a hazardous air pollutant in U.S. Environmental Pretection Agency (EPA) regulations, highlighting its toxicity. In recent years, there has been a significant rise in global warming potential worldwide, which is largely driven by increased energy demands by chloromethane industries. Therefore, estimating the energy demands of the process routes of the chloromethane plants is crucial to overcome the complexities involved in the chloromethane plants. Furthermore, the annual production of methyl chloride is approximately 264 million pounds for a plant. Because of its carcinogenic nature and the potential for spills, there is an urgent requirement to identify mathematical diagnostic systems to assess and manage its toxicity, based on plant data. The upstream processes of the methyl chloride industry involve multiple reactors and separation equipment. A detailed process flow diagram is given in Figure 1. The reactions involved in the production process are provided in Equations (1) and (2).

Developing a reliable method to calculate the energy and exergy framework for industrial-scale methyl chloride processes based on the above reactions could enhance daily operational efficiency. However, determining the energy recovery rate for this process is complex. It requires the evaluation of thermodynamic losses due to the heat of reaction from reactant and product species given in the equations. Reactant species involved in methyl chloride formation contribute significantly to these heat losses compared to other process components. Consequently, the reactor in the process may experience higher enthalpy or entropy loss during the production process. A theoretical framework to evaluate the conceptual reactor and validate it using real-time data could be the first step to creating the diagnostic tool for optimizing the chloromethane process.

$$CH_3OH + HCl \rightarrow CH_3Cl + H_2O \quad (1)$$

$$2CH_3OH \rightarrow CH_3OCH_3 \quad (2)$$

Soft computing and data science models provide significant advantages over traditional regression methods for calculating the energy and exergy of the chloromethane process. Data science models reported for similar processes show flexibility, speed, accuracy, and applicability within the constrained timeframes in developing self-learning predictive maintenance models (Dobbelaere et al., 2021). Among the various models documented in the literature, machine learning (ML) has demonstrated high accuracy in predicting thermo-dynamic properties and reporting reactor failures, ML models are aiding in the estimation of reactor performance indices (López-Guajardo et al., 2021). However, obtaining a real-time dataset for a year or month from the chloromethane plant remains a challenging task.

* Corresponding author. Tel.:91-9626845903.
E-mail address: shankarraman.d@vit.ac.in (S. R. Dhanushkodi).

ISSN: 2663-6859 print/2663-6867 online
 doi:10.3808/jeil.202400142

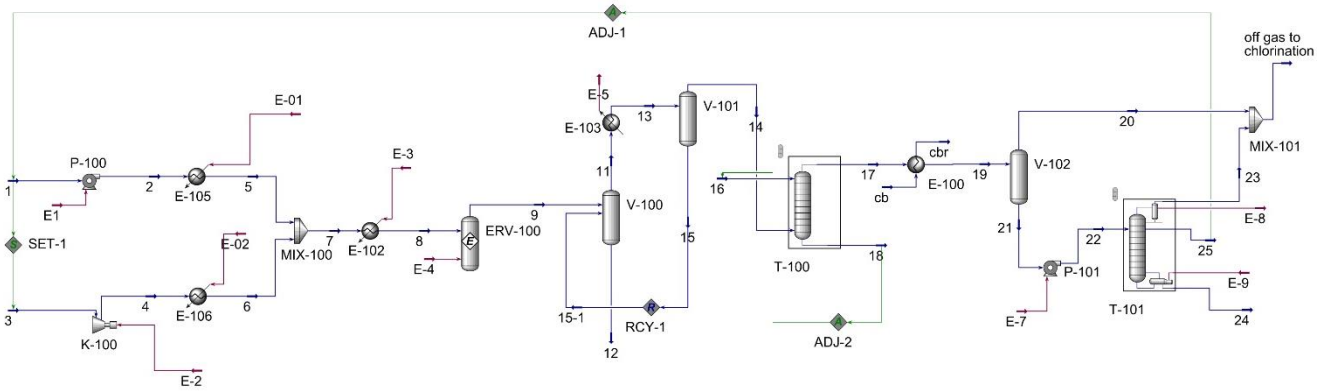

Figure 1. The process flow diagram of methyl chloride using hydrochlorination of methanol.

Table 1. Literature Finding in HYSYS Models

No.	Literature	Reference
1	A conceptual ethylene plant is simulated and operating parameters and the error variation among the different machine learningmodels are identified	Zhu et al. (2018)
2	The ethylene and propene plant are simulated, and the accuracy and performance of process is evaluated accurately by the machine learningmodel	Zhu et al. (2020)
3	Forecast of shale gas process reservoirs is performed using SVR, ANN, RNN, and GPR algorithms and process precision is calculated	Syed et al. (2022)
4	The exergy analysis of the synthetic natural gas process is simulated using the feed rate and pressure differences	Qadeer et al. (2022)
5	The backpropagation neural network algorithm has been identified as a promising model for predicting thermal properties in waste heat systems	Wang et al. (2020)
6	Twelve different algorithms were applied to generate ANN topology for accurate predictions in methane production from waste	Ozarslan et al. (2021)
7	ANN model is developed to assess the efficiency of the biogas production behavior	Ghatak and Ghatak (2018)
8	Chemical looping oxygen carrier process parameters are assessed by ANN to reduce extra experimental work and testing methods of evaluation	Yan et al. (2020)
9	ANN is applied to predict the waste heat energy from the ORC system	Yang et al. (2018)
10	The dryer's energy loss and exergy destruction input parameters were used in the ANN model with 0.995 and 0.996 prediction accuracies	Kaveh et al. (2021)
11	ANN is used to evaluate the exergy of syngas based on the Aspen Plus experimental conditions	Sezer et al. (2021)
12	Semi-supervised models in production lines to enhance the knowledge behind the soft computing technique	Kang et al. (2020)
13	Supervised data-driven GB and ANN models reported to predict the relation between the input and output datasets of drainage recovery process reservoir.	Huang and Chen (2021)

Shang et al. (2018) utilized Aspen Hyprotech Systems (HYSYS) to generate datasets for various processes for chemical industries and employed ML models to optimize reactor operations. Their model analyzes the transport parameters of the process. Javad (2021) used ML to predict energy and exergy in assessing the Kalina cycle. They applied supervised ML models to forecast hydrogen yield in dark fermentation processes. Additionally, Hosseinzadeh et al. (2022) utilized Random Forest regression (RFR) and Adaboost Regression (ABR) to evaluate process parameters and assess the efficiency of wastewater treatment. In another study, Fang et al. (2021) applied K Nearest Neighbor (KNN), Decision Tree Regression (DTR), Support Vector Regression (SVR), and Artificial Neural Network (ANN) models to simulate and evaluate the dehydrogenation process. Cai et al. (2021) utilized Recurrent Neural Network models to analyze fixed bed reactors and extract thermodynamic and kinetic process parameters. Sezer and Özveren (2021) simulated the Gibbs reactor using Aspen Plus to develop an ANN model to predict exergy values and efficiency. Bhadriraju et al. (2019) applied ML models to assess nonlinearity in continuous reactors for process optimization. A summary of the applications of different machine-learning models can be found in Table 1. A similar framework is reported by Zhao et al. (2021), Mageed (2021) and Sultana et al. (2022). They have used these concepts to evaluate the performances of the biomass gasification processes. Zhang et al. (2021) demonstrated a Gradient Boosting Regression (GBR) framework to optimize the plants in chloromethane production unit. However, descriptions of data, and validation of the model with the exergy or energy real-time data are not discussed.

Several supervised machine-learning models reported in the literature have shown advancements in predicting energy loss

and exergy destruction in chemical processes. However, there has been no study that aims to utilize the HYSYS-ML framework to investigate the hydrochlorination of the methanol process for methyl chloride production. This process requires the estimation of residual errors and the identification of a suitable ML model to address the inherent complexity and variability in the process (Sultana et al., 2022). Therefore, we aim to bridge this research gap by employing soft computing models with HYSYS to predict the performance of the methyl chloride process. Based on the literature gap, we have defined clear objectives to assess the chloromethane process using the hydrochlorination pathway, illustrated in Figure 2. They are:

(1) Formulating an HYSYS model to generate a methyl chloride plant dataset to understand the parameters and to set up a model based on a plant that needs to be investigated.

(2) Assessing the competencies of the HYSYS model and generate the plant datasets using hyperparameters.

(3) Validating the competencies of the simulated data set and train it with supervising data science model available in the literature.

(4) Comparing all data science models to identify the competent model to develop the HYSYS-ML model framework.

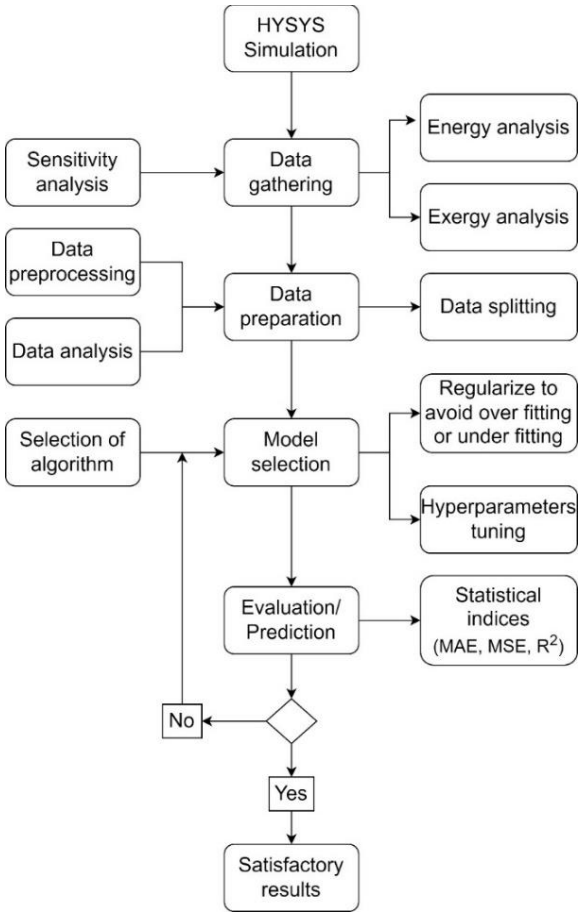

Figure 2. Flow chart of ML-based loss prediction in methyl chloride reactor.

This framework introduces a novel approach aimed at enhancing decision-making process through predictive maintenance models based on ML for the chloromethane plant. Overall methods adopted to solve the objective are given in Figure 2. Integration of HYSYS simulation with ML models improves the capabilities of traditional process simulation tools, yields accurate predictions, performs accurate optimizations, and helps decision-making during plant operation. The framework could predict equipment failures, provide strategies for real-time adjustments and provide suggestions based on different operating conditions.

2. Methodology

The research objectives will be executed using the following steps: (1) Gather data for the methyl chloride process using the HYSYS reactor simulation tool (Section 2.1); (2) Assess the focus on hyperparameter optimization (Sections 2.1, 2.2 and 2.3); (3) Compare and contrast various supervised ML models (Section 2.4); (4) Analyze residual errors to determine the most effective ML model in the process (Section 3.2).

2.1. Reactor Simulation

A design for a Methyl chloride plant and a detailed balance of plant information is integrated into the process flow diagram (PFD) (Figure 2). This PFD is then exported to the HYSYS simulator to extract raw data pertinent to ML models. The simulator tools are used to process both thermodynamic and kinetic data for the reactor components. The property estimator placed in the simulation tool is used to extract and refine this kinetic and thermodynamic data. Mass and thermal energy balances are computed using Equations (3) ~ (5). A detailed methodology outlined by Gollangi and NagamalleswaraRao (2022) in our previous work is adopted to analyze the heat flow, energy degradation, and exergy destruction within the process. The fundamental equations for these computations are provided below:

The mass balance equation for the total process is:

$$m_{in} = m_{out} \tag{3}$$

where m_{in} and m_{out} are mass in and mass out. The thermal energy equilibrium equation is as:

$$E_{in} - E_{out} + Q_{in} - Work_{out} = E_l \tag{4}$$

where E_{in}, E_{out}, and E_l are energies for the inlet, outlet and system respectively. Q_{in} is the heat input and $Work_{out}$ is the output work. The exergy balance equation for a system is:

$$Ex_{in} - Ex_{out} + Ex_{heat} - Ex_{work} = I_d \tag{5}$$

where Ex_{in} and Ex_{out} are exergies for the inlet and outlet respectively. Ex_{heat} and Ex_{work} are exergies due to heat and work respectively. I_d is the irreversibility. As higher exergy destruction is observed in the real-time plant, all calculations used in this

study account for irreversibility in the reactions. All calculations are conducted for the following parameters molar feed rate, reaction, temperature, specific energy input, and specific exergy. Each parameter is evaluated using three different datasets. The parameters contributing significantly to losses were identified, and a sensitivity analysis was performed following the procedure outlined by Yandrapu and Kanidarapu (2021). We use initial datasets from the HYSYS tool to perform the analysis as this step reduces computation time and costs. Results include tabulated data for both energy loss and exergy destruction given in Table 2.

Table 2. Input and Output Parameters and Ranges of ML Model

Parameter Type	Parameters	Range
Input Variable	Molar Feed Rate (kg/hr)	1151.70 ~ 1413.50
	Reactor Temperature (K)	350.00 ~ 450.00
	Energy Input (kW)	12302.48 ~ 15098.51
	Exergy Input (kW)	625.60 ~ 767.78
Output Variable	Energy Loss (kW)	9028.60 ~ 13085.20
	Energy Destruction (kW)	247.17 ~ 375.86

2.2. Data Collection

Gollangi and NagamalleswaraRao (2022) reported the importance of discerning interaction effects in assessing energy and exergy losses in chemical plant data collection. Therefore, to obtain additional data on thermodynamic losses, a sensitivity analysis was conducted. The molar feed rate of the reactant and reaction temperature were varied, yielding multiple datasets for ML models. A total of 235 experimental conditions were selected, and energy balances were calculated for each. All simulated and experimental data points were normalized using a standard normalization factor within the range of –1 to 1. This approach ensures that the dataset avoids overfitting, underfitting, and complexity.

2.3. Data Treatment

HYSYS datasets were pre-processed using supervised ML data-driven models. All datasets are classified to detect and handle missing values, while nonlinearity in the data points is addressed through feature scaling. A total of 235 datasets are utilized in this study. These are divided into training and testing sets as detailed in the following section: 188 datasets are used for training, and 47 datasets are reserved for testing purposes.

2.4. ML Model Selection

Selecting the most suitable ML model is critical for the methyl chloride industry. The model should strike a balance between simplicity and complexity and must be tailored to achieve optimal performance. Eight prominent supervised ML regression techniques were developed to evaluate the dataset generated by the Aspen HYSYS tool. These include Hyperparameters of these models were fine-tuned using data extracted from the Scikit-learn library, enabling a comprehensive analysis of the chloromethane plant. The neural network topology used in this study is illustrated in Figure 3. All models comprise four input parameters: molar feed rate (*mf*), reactor temperature (*T*), energy input (*En*), and exergy input (*ex*) in the input layer. It includes a hidden layerwith appropriate neurons and weights, and an output layer with two parameters. The ranges of input and output variables utilized to construct the HYSYS-ML framework are detailed in Table 2. Training and testing data are split in a ratio of 80% for training and 20% for testing.

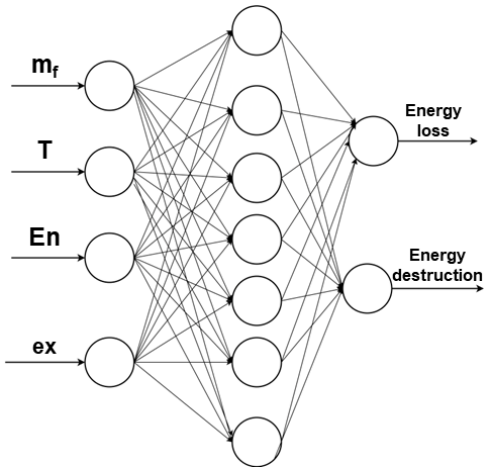

Figure 3. The topological structure for the developed ML neural network (m_f is mass flow rate; *T* is temperature; *En* is energy input and *ex* is exergy input).

2.5. Statistical Analysis

We performed statistical analysis on all eight ML models, assessing critical metrics including *R*-squared (R^2), mean absolute error (MAE), and mean squared error (MSE) as defined in Equations (6) ~ (8) (Geng et al., 2018; Budamala and Baburao Mahindrakar, 2019). The goal is to identify the most effective ML model for precise predictions of energy loss and exergy destruction within the plant:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{pr,i} - y_{tr,i})^2 \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{pr,i} - y_{tr,i})}{\sum_{i=1}^{N}(y_{pr,i} - y_m)} \tag{7}$$

$$MAE = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)}{N} \tag{8}$$

where, y_i x_i and N are the predicted values, observed values, and the number of data points, respectively; $y_{pr,i}$ and $y_{tr,i}$ predicted and true values of methyl chloride process; y_m the average of true methyl chloride process values.

3. Results and Discussion

3.1. Learning Curve

A learning curve reported in Figure 4 is used as a benchmark tool to evaluate the ML algorithm employed in the study. While all models were tested using training and testing datasets, the detailed results for SVR are presented in the Figure 4. The graph illustrates the optimal learning rate based on MSE versus epochs, where higher epochs correspond to longer computation times for training and testing by updating neuron weights. A significant reduction in error value is observed at the 80th iteration. It indicates improved validation results for the models. The accuracy of SVR is further assessed through the learning curve, revealing no signs of overfitting or underfitting in the data analysis—all models demonstrate a good fit. The trend line equation confirms minimal deviation from data points. While the graph shows a small set of input data points used for testing the model and creating the learning curve, both datasets demonstrate optimal values for the loss function across identical parameters. The learning curve of this model shows no signs of bias or variance errors, ensuring precise predictions for datasets of varying sizes without any erroneous outputs. Figure 4 serves as a reliable tool to validate the models and good fit confirms the absence of underfitting or overfitting in the proposed models.

3.2. Inference of Sensitivity Analysis

We conducted a precise sensitivity analysis on reactor unit operations using HYSYS for the hydrochlorination of the methanol process under consistent operating conditions. The reaction temperature emerged as a critical parameter that can significantly impact thermodynamic losses due to its association with an exothermic chemical reaction. Higher reaction temperatures may enhance the yield rate of the desired product but also lead to increased losses. Additionally, the feed rate to the reactor, as described in the equation, influences these losses. Moreover, substantial energy and exergy input in the form of heat duty also plays a pivotal role in these losses.

3.3. Comparison of Models

This section reports the predictive maintenance capabilities of each model along with their respective statistical details.

3.3.1. Bayesian Ridge Regression

To enhance prediction accuracy in the multioutput Bayesian Ridge Regression (BRR) model, we fine-tuned hyperparameters individually for each output using grid search. We employed a linear kernel function with default values for alpha and lambda to estimate precision and weights. In the training phase, the model achieved an R^2 of 0.9982 for predicted energy loss, with MSE and MAE values of 0.0009 and 0.0243, respectively. For exergy destruction, the training phase yielded the values of R^2, MSE, and MAE at 0.9801, 0.0009, and 0.0265, while in testing, these metrics improved to 0.9982, 0.0001, and 0.0070. Overall, the BRR multioutput model demonstrated satisfactory performance for the process.

3.3.2. AdaBoost Regression

The AdaBoost multioutput regressor model was optimized using grid search, employing a linear kernel and a learning rate of 1.0. It utilized 50 n_estimators to compute indices for both energy loss and exergy destruction. During training, the model achieved MSE (0.0017 and 0.0042, R^2 (0.9646 and 0.9178), and MAE (0.0342 and 0.0517) for energy loss prediction. For exergy destruction, the testing and training phases yielded MSE (0.0026 and 0.0040), R^2 (0.9450 and 0.9203), and MAE (0.0405 and 0.0547), respectively. These findings, elaborated in the Tables 3 and 4, were utilized to evaluate the performance of the reactor process and gauge the accuracy of predictions for both energy loss andexergy destruction parameters.

3.3.3. Gradient Boosting Regression

We fine-tuned hyperparameters of the GBR model to optimize its performance in predicting loss functions. This included adjusting the number of boosting stages, maximum depth of

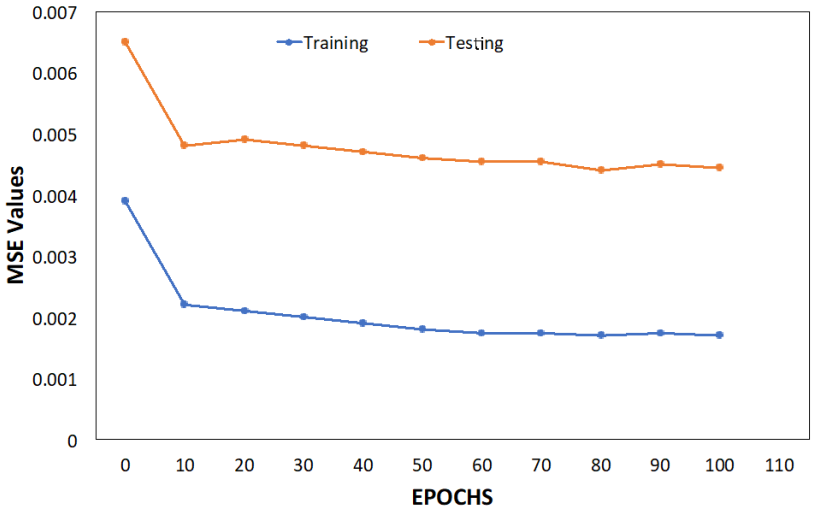

Figure 4. Learning rate curve between training and testing MSE values.

Table 3. Performance Comparison of All ML Models in Energy and Exergy Loss Prediction

Energy Loss Prediction						
ML Model	Training			Testing		
	MAE	MSE	R^2	MAE	MSE	R^2
ABR	0.0342	0.0019	0.9646	0.0517	0.0042	0.9178
ANN	0.0410	0.0025	0.9467	0.0425	0.0032	0.9367
BRR	0.0070	0.0000	0.9982	0.0243	0.0009	0.9815
GBR	0.0739	0.0087	0.8175	0.0895	0.0122	0.7585
KNR	0.0080	0.0002	0.9968	0.0242	0.0009	0.9817
RFR	0.0088	0.0001	0.9973	0.0313	0.0014	0.9724
SGD	0.0109	0.0002	0.9966	0.0251	0.0009	0.9823
SVR	0.0338	0.0020	0.9588	0.0428	0.0033	0.9344
Exergy Loss Prediction						
ML Model	Training			Testing		
	MAE	MSE	R^2	MAE	MSE	R^2
ABR	0.0405	0.0026	0.9450	0.0547	0.0040	0.9203
ANN	0.0593	0.0054	0.8868	0.0592	0.0049	0.9027
BRR	0.0265	0.0009	0.9801	0.0070	0.0001	0.9982
GBR	0.0776	0.0093	0.8033	0.0899	0.0119	0.7639
KNR	0.0245	0.0009	0.9810	0.0148	0.0003	0.9939
RFR	0.0248	0.0009	0.9802	0.0185	0.0005	0.9899
SGD	0.0284	0.0010	0.9782	0.0107	0.0002	0.9963
SVR	0.0276	0.0014	0.9714	0.0326	0.0019	0.9615

Table 4. Performance Comparison of All ML Models

Model	Interpretability	Computational Complexity	Robust to Outliers	Scalable for Plant	Merit	Demerit
BRR	YES	NO	NO	YES	Probabilistic Predictions	Assumes Linear Cases
ANN	YES	YES	YES	YES	Predict Non-linear Relationships	Big Datasets Needed
GBR	YES	YES	YES	YES	Robust	Overfitting
KNR	YES	YES	NO	NO	Easy to Adopt	Computational Cost
RFR	NO	YES	YES	NO	Needs High-Dimensional Data	Hyper Parameter Tuning
SGD	YES	NO	YES	YES	NeedS Large Datasets,	Sensitive
SVR	YES	YES	YES	NO	Needs High-Dimensional Data	Cost

GBR trees, and minimum samples required to split internal and leaf nodes (set to 100, 3, 2, and 1, respectively). During training and cross-validation of datasets, we obtained the following metrics for energy loss prediction: R^2 values of 0.8175 and 0.7585, MAE values of 0.0739 and 0.0895, and MSE values of 0.0087 and 0.0122 Similarly, for exergy destruction, the metrics were R^2 values of 0.8033 and 0.7639, MAE values of 0.0899 and 0.0776, and MSE values of 0.0093 and 0.0119.

3.3.4. Neural Network Regressor

The MLP Regression ML model was developed with fine-tuned hyper-parameters including 100 neurons in a hidden layer, the Adam solver optimized for large datasets, RELU activation function, and default epsilon value for numerical stability. Training and testing the model on the SVR yielded the following indices for energy loss: R^2 (0.9467, 0.9367), MSE (0.0025, 0.0032), and MAE (0.0410, 0.0425). Similarly, for exergy destruction, the results were R^2 (0.9026, 0.8868), MSE (0.0049, 0.0054), and MAE (0.0591, 0.0593) in training and testing, respectively. Table 3 illustrates the fitting and correlation between true and predicted values of energy loss and exergy destruction in our model, confirming its alignment with plant data.

3.3.5. Nearest Neighbors Regression

The K-neighbors regressor is a supervised machine-learning model that predicts targets by interpolating from the nearest neighbors in training datasets. We configured a practical model with optimized hyperparameters, including selecting five neighbors based on Euclidean distance with uniform weight across all neighbourhood training datasets, and a leaf size of 30 for improved construction speed. This multioutput regressor calculated statistical errors during the training phase for energy loss and exergy destruction, as well as during validation for both output parameters, detailed in Tables 3 and 4. The model's bestfit line and its predictive accuracy with true values for energy loss and exergy destruction in reactor operation align well with simulated plant data.

3.3.6. Random Forest Regression

The RF regressor is a supervised ensemble ML algorithm

puts from multiple ML algorithms. For this model, hyperparameters were fine-tuned to values of 100 boosted trees in the forest, minimum samples required for split set to 2, leaf node samples requirement of 1, and the feature used for the best split was seven. Figure 5 illustrate the prediction accuracy and model advantages using testing values compared to actual data points. Statistical error indices for energy loss (e-loss) and exergy destruction (ex-des) were computed for both the training and testing phases.

3.3.7. Stochastic Gradient Descent Regression

The SGD regressor was built with hyperparameters fine-tuned via grid search, setting the maximum iterations of fitting to 1000, using l2 regularization with a penalty of 0.001 to avoid partial fitting, and adjusting epsilon to enhance the learning rate of the SGD model. Table 3 displays a plot illustrating the validation trend and strength variations of the model compared to actual values. The error indices R^2 (0.9966 and 0.9823), MSE (0.00016 and 0.0009), and MAE (0.0109 and 0.0251) were computed for energy losses based on observed and predicted data. Similarly, for exergy destruction, the model yielded R^2 (0.9781 and 0.9963), MSE (0.00103 and 0.0002), and MAE (0.0284 and 0.0107). These metrics showcase the model's effectiveness in predicting both parameters.

3.3.8. Support Vector Regression

The SVR aims to find the optimal fitting line in extensive-datasets, encompassing as many data points as possible. Various

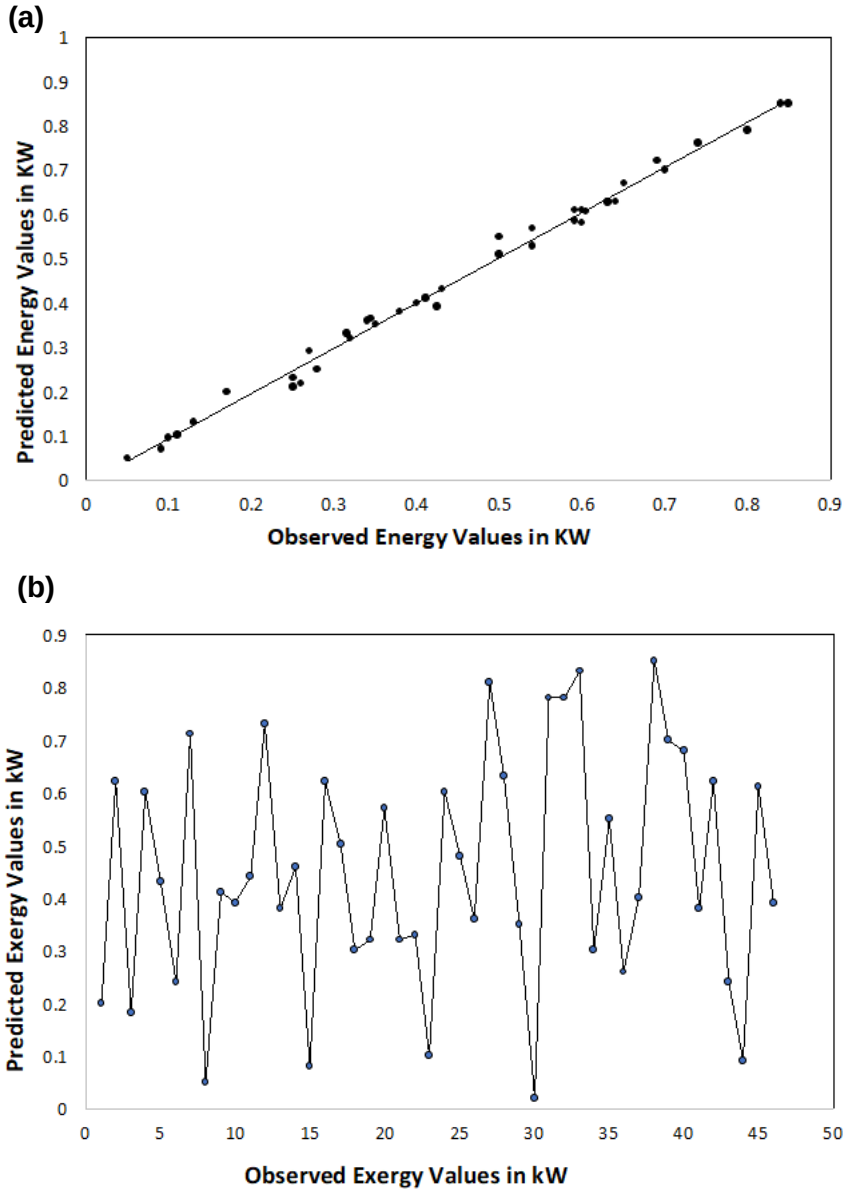

Figure 5. Correlation coefficient and prediction strength for energy and exergy losses predicted by the BRR model.

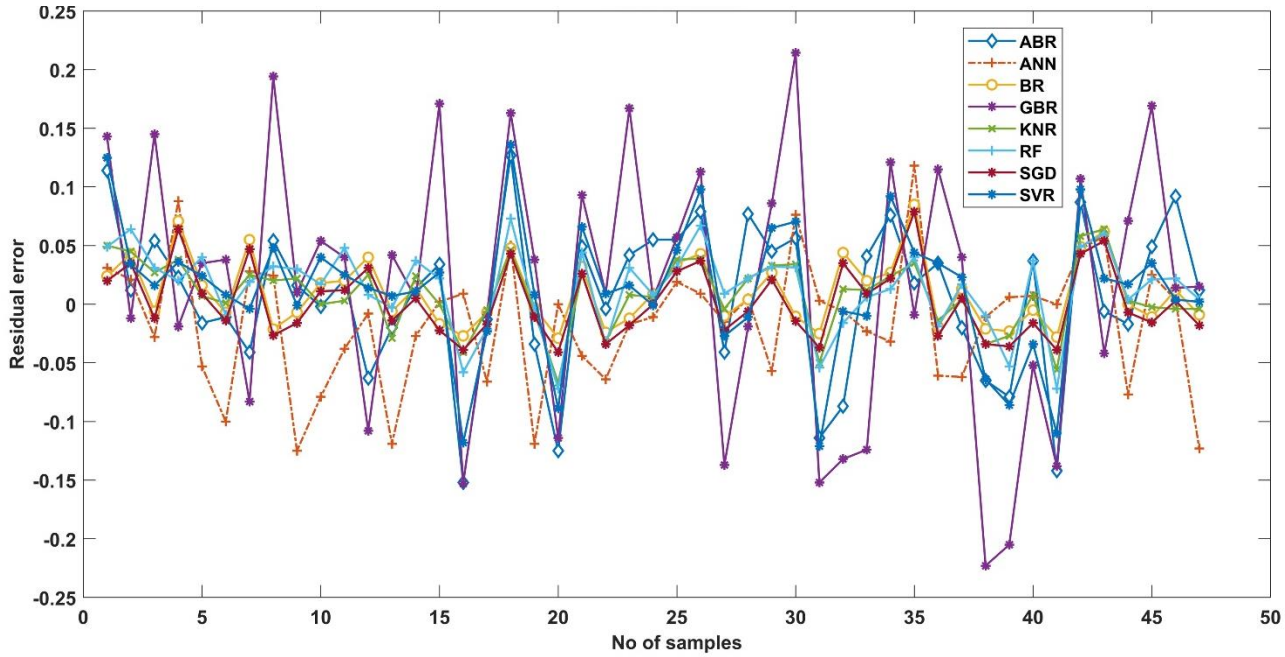

Figure 6. Comparison based on residual errors calculated.

conditions of the SVR model's hyper-parameters are fine-tuned using grid search to determine optimal statistical metrics. Among these, the RBF kernel function demonstrates optimal parameters with C, epsilon, and degree values of 1, 0.1, and 3, respectively. Performance metrics such as R^2, MAE, and MSE are then computed from the model to evaluate its accuracy and effectiveness.

3.3.9. Comparison of ML models

The predictive performance of all models is evaluated based on residual errors during testing. Figure 6 illustrates the residual error charts for eight models predicting exergy destruction in the reactor. Notably, the BRR model exhibits dispersed residuals close to the zero line, indicating superior predictive capability compared to other models. Additionally, prediction results from ML models are assessed using statistical metrics R^2, MSE, and MAE. Tables 3 and 4 present these indicators for evaluating energy loss and exergy destruction in a methyl chloride reactor. Overall, all models perform well, showing comparable R^2 values and nearly identical MSE calculations. Notably, BRR, SGD, KNR, and RFR achieve notable performance in predicting thermodynamic losses, with high R^2 values (0.998, 0.996, 0.993, and 0.989, respectively), low MSE values ($9.2e^{-5}$, $1.8e^{-4}$, $3.1e^{-4}$, and $5.1e^{-4}$ respectively), and minimal MAE values (0.007, 0.011, 0.014, and 0.0185 respectively). Following them, SVR, ABR, ANN, and GBR also demonstrate accurate predictions inassessing reactor process performance. Choosing the correct one for the HYSYS-ML framework is cumbersome since the choice of the model depends on various factors. By experimenting with different models, we performed crossvalidation to determine the best model for the chloromethane plant. We chose interpretability needs and computational resources as two important criteria. The results are reported in Table 4. Therefore, BRR is preferred over other models owing to low cost and computer complexity.

3.4. Applicability of the Framework

Combining the HYSYS simulation tool with ML models improve the model accuracy. It can solve nonlinearities in the process dynamics and assess the complex interactions with the data sets and the tools. The framework can provide faster iterative methods than the traditional ones. It can predict equipment failures; and plant maintenance issues and provides control suggestions using both the real-time operational data and ML model output. Furthermore, it will be useful to identify the risks associated with the operation and toxicity level based on the plant data. Our framework can provide adaptive control strategies based on feedback from the framework and process data. It can simulate various operating conditions on HYSYS outputs and historical data predict the risks based on the simulations, however showcasing all applicability is beyond the scope of the present study.

4. Conclusions

Our primary objective in this study is to simulate the methyl chloride production process using Aspen HYSYS software and utilize the extracted dataset to select an appropriate ML model. Our framework integrates HYSYS simulator data with ML models, aiming to optimize and evaluate methyl chloride production via methanol hydrochlorination. This approach introduces several novel aspects to our research. Here are the key findings emphasized in our work:

(1) Dataset Generation and Sensitivity Analysis: We generated datasets using Aspen HYSYS simulations for the methyl

chloride process. These datasets serve as a comprehensive database to assess and optimize production parameters. Sensitivity analysis was conducted on the selected reactor, and input parameters were generated to train machine-learning models.

(2) ML Model Selection: We identified several ML models suitable for the dataset generated by the simulator. These include BRR, ABR, ANN, GBR, KNR, RFR, SGD, and SVR.

(3) Statistical Evaluation: Statistical indices were computed to assess the performance of all machine-learning models in predicting the chloromethane process. Hyperparameter tuning and residual error analysis were used to identify the most suitable ML model for the process.

(4) Performance of ML Models: All ML models demonstrated superior prediction capabilities during both training and testing phases. For energy loss prediction, SGD, KNR, BRR, and RFR achieved low errors of 0.00090, 0.00092, 0.00093, and 0.00139 respectively, with corresponding R^2 values of 0.9822, 0.9817, 0.9815, and 0.9723. As for predicting exergy destruction, BRR, SGD, KNR, and RFR achieved low error values of 0.00009, 0.00019, 0.00031, and 0.00051 respectively, with corresponding R^2 values of 0.9981, 0.9962, 0.9939, and 0.9981. Notably, BRR consistently showed a superior fit within the HYSYS-ML framework compared to all other models.

In summary, our integrated approach of using HYSYS simulations and ML models offers a robust framework for optimizing and evaluating methyl chloride production processes. The findings underscore the effectiveness of BRR in the developed framework, particularly in predicting exergy destruction.

References

Bhadriraju, B., Narasingam, A. and Kwon, J. (2019). Machine learning-based adaptive model identification of systems: Application to a chemical process. *Chemical Engineering Research and Design*, 152, 372-383. https://doi.org/10.1016/J.CHERD.2019.09.009

Budamala, V. and Baburao Mahindrakar, A. (2019). Enhance the prediction of complex hydrological models by pseudo-simulators. *Geocarto International*, 36(9), 1027-1043. https://doi.org/10.1080/10106049.2019.1629646

Cai, Q., Luo, X., Gao, C., Guo, P., Sun, S., Yan, S. and Zhao, P. (2021). A machine learning-based model predictive control method for pumped storage systems. *Frontiers in Energy Research*, 9, 716. https://doi.org/10.3389/FENRG.2021.757507/BIBTEX

Dobbelaere, M.R., Plehiers, P.P., van de Vijver, R., Stevens, C. V. and van Geem, K. M. (2021). Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. *Engineering*, 7(9), 1201-1211. https://doi.org/10.1016/J.ENG.2021.03. 019

Fang, H., Zhou, J., Wang, Z., Qiu, Z., Sun, Y., Lin, Y., Chen, K., Zhou, X. and Pan, M. (2021). Hybrid method integrating machine learning and particle swarm optimization for smart chemical process operations. *Frontiers of Chemical Science and Engineering*, 16(2), 274-287. https://doi.org/10.1007/S11705-021-2043-0

Geng, Z., Li, H., Zhu, Q. and Han, Y. (2018). Production prediction and energy-saving model based on Extreme Learning Machine integrated ISM-AHP: Application in complex chemical processes. *Energy*, 160, 898-909. https://doi.org/10.1016/j.energy.2018.07.077

Ghatak, M.D. and Ghatak, A. (2018). Artificial neural network model to predict behaviours of biogas production curve from mixed lignocellulosic co-substrates. *Fuel*, 232, 178-189. https://doi.org/10.1016/J.FUEL.2018.05.051

Gollangi, R. and NagamalleswaraRao, R. (2022). Energy, exergy analysis of conceptually designed monochloromethane production process from hydrochlorination of methanol. *Energy*, 239, 121858. https://doi.org/10.1016/J.ENERGY.2021.121858

Hosseinzadeh, A., Zhou, J.L., Altaee, A. and Li, D. (2022). Machine learning modelling and analysis of biohydrogen production from wastewater by dark fermentation process. *Bioresource Technology*, 343, 126111. https://doi.org/10.1016/J.BIORTECH.2021.126111

Huang, Z. and Chen, Z. (2021). Comparison of different machine learning algorithms for predicting the SAGD production performance. *Journal of Petroleum Science and Engineering*, 202, 108559. https://doi.org/10.1016/J.PETROL.2021.108559

Javad Dehghani, M. (2021). Enhancing energo-exergo-economic performance of Kalina cycle for low- to high-grade waste heat recovery: Design and optimization through deep learning methods. *Applied Thermal Engineering*, 195, 117221. https://doi.org/10.1016/J.APPLTHERMALENG.2021.117221

Kang, Z., Catal, C. and Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering*, 149, 106773. https://doi.org/10.1016/J.CIE.2020.106773

Kaveh, M., Chayjan, R.A., Golpour, I., Poncet, S., Seirafi, F. and Khezri, B. (2021). Evaluation of exergy performance and onion drying properties in a multi-stage semi-industrial continuous dryer: Artificial neural networks (ANNs) and ANFIS models. *Food and Bioproducts Processing*, 127, 58-76. https://doi.org/10.1016/J.FBP. 2021.02.010

López-Guajardo, E.A., Delgado-Licona, F., Álvarez, A. J., Nigam, K. D. P., Montesinos-Castellanos, A. and Morales-Menendez, R. (2021). Process intensification 4.0: A new approach for attaining new, sustainable and circular processes enabled by machine learning. *Chemical Engineering and Processing-Process Intensification*, 108, 108671. https://doi.org/10.1016/J.CEP.2021.108671

Mageed, A.K. (2021). Modeling photocatalytic hydrogen production from ethanol over copper oxide nanoparticles: a comparative analysis of various machine learning techniques. *Biomass Conversion and Biorefinery*, 13, 3319-3327. https://doi.org/10.1007/S13399-021-01388-Y

Özarslan, S., Abut, S., Atelge, M.R., Kaya, M. and Unalan, S. (2021). Modeling and simulation of co-digestion performance with artificial neural network for prediction of methane production from tea factory waste with co-substrate of spent tea waste. *Fuel*, 306, 121715. https://doi.org/10.1016/J.FUEL.2021.121715

Qadeer, K., Ahmad, A., Naquash, A., Qyyum, M. A., Majeed, K., Zhou, Z., He, T., Nizami, A.S. and Lee, M. (2022). Neural network-inspired performance enhancement of synthetic natural gas liquefaction plant with different minimum approach temperatures. *Fuel*, 308, 121858. https://doi.org/10.1016/J.FUEL.2021.121858

Sezer, S., Kartal, F. and Özveren, U. (2021). Prediction of chemical exergy of syngas from downdraft gasifier by means of machine learning. *Thermal Science and Engineering Progress*, 26, 101031. https://doi.org/10.1016/J.TSEP.2021.101031

Sezer, S. and Özveren, U. (2021). Investigation of syngas exergy value and hydrogen concentration in syngas from biomass gasification in a bubbling fluidized bed gasifier by using machine learning. *International Journal of Hydrogen Energy*, 46(39), 20377-20396. https://doi.org/10.1016/J.IJHYDENE.2021.03.184

Shang, J., Gu, X., Yang, L., Tang, H., Zhang, K. and Ji, Z. (2018). Preference-driven yield-and-quality optimization for high-sulfur gas sweetening process by extreme learning machine model. *Cluster Computing*, 22(3), 6371-6381. https://doi.org/10.1007/S10586-01 8-2136-9

Sultana, N., Hossain, S. M.Z., Abusaad, M., Alanbar, N., Senan, Y. and Razzak, S. A. (2022). Prediction of biodiesel production from microalgal oil using Bayesian optimization algorithm-based machine learning approaches. *Fuel*, 309, 122184. https://doi.org/10.1016/J.F

UEL.2021.122184

Syed, F.I., Alnaqbi, S., Muther, T., Dahaghi, A.K. and Negahban, S. (2022). Smart shale gas production performance analysis using machine learning applications. *Petroleum Research*, 7(1), 21-31. https://doi.org/10.1016/J.PTLRS.2021.06.003

Wang, J., Zhai, Y., Yao, P., Ma, M. and Wang, H. (2020). Established prediction models of thermal conductivity of hybrid nanofluids based on artificial neural network (ANN) models in waste heat system. *International Communications in Heat and Mass Transfer*, 110, 104444. https://doi.org/10.1016/J.ICHEATMASSTRANSFER.2019.104444

Yan, Y., Mattisson, T., Moldenhauer, P., Anthony, E.J. and Clough, P.T. (2020). Applying machine learning algorithms in estimating the performance of heterogeneous, multi-component materials as oxygen carriers for chemical-looping processes. *Chemical Engineering Journal*, 387, 124072. https://doi.org/10.1016/J.CEJ.2020.124072

Yandrapu, V.P. and Kanidarapu, N.R. (2021). Process design for energy efficient, economically feasible, environmentally safe methyl chloride production process plant: Chlorination of methane route. *Process Safety and Environmental Protection*, 154, 360-371. https://doi.org/10.1016/J.PSEP.2021.08.027

Yang, F., Cho, H., Zhang, H., Zhang, J. and Wu, Y. (2018). Artificial neural network (ANN) based prediction and optimization of an organic Rankine cycle (ORC) for diesel engine waste heat recovery. *Energy Conversion and Management*, 164, 15-26. https://doi.org/10.1016/J.ENCONMAN.2018.02.062

Zhang, W., Li, J., Liu, T., Leng, S., Yang, L., Peng, H., Jiang, S., Zhou, W., Leng, L. and Li, H. (2021). Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. *Bioresource Technology*, 342, 126011. https://doi.org/10.1016/J.BIORTECH.2021.126011

Zhao, S., Li, J., Chen, C., Yan, B., Tao, J. and Chen, G. (2021). Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. *Journal of Cleaner Production*, 316, 128244. https://doi.org/10.1016/J.JCLEPRO.2021.128244

Zhu, Q.X., Wang, X., He, Y.L. and Xu, Y. (2018). An improved extreme learning machine integrated with nonlinear principal components and its application to modelling complex chemical processes. *Applied Thermal Engineering*, 130, 745-753. https://doi.org/10.1016/J.APPLTHERMALENG.2017.11.061

Zhu, W., Liu, X., Hou, X., Hu, J. and Diao, Z. (2020). Application of machine learning to process simulation of n-pentane cracking to produce ethylene and propene. *Chinese Journal of Chemical Engineering*, 28(7), 1832-1839. https://doi.org/10.1016/J.CJCHE.202 0.01.01