Available *online* at : http://jnte.ft.unand.ac.id/

Jurnal Nasional Teknik Elektro

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

Predictive Modeling of Carbon Monoxide with MOS Sensors and Machine Learning: A Potential Tool for Process Safety Improvement

Hermin Kartika Sari [1], Thomas Oka Pratama [2], Yohana Fransiska Ferawati [1], Gita Nur Sajida [1], Gustin Mustika Krista [1], Teguh Taufiqurohim [1], Shoerya Shoelarta [1]*

[1] *Department of Chemical Engineering, Politeknik Negeri Bandung, Indonesia*
[2] *Department of Engineering Physics and Nuclear Engineering, Universitas Gadjah Mada*

ARTICLE INFORMATION

Received: July 06, 2025
Revised: March 12, 2026
Accepted: March 27., 2026
Available online: March 29, 2026

KEYWORDS

CO Detection, MOS Sensor, Feature Selection, Machine Learning, Process Safety Monitoring

CORRESPONDENCE

Phone: +62 85743272745
E-mail: hermin.kartika@polban.ac.id

A B S T R A C T

Carbon monoxide (CO) is a toxic, odorless gas commonly present in industrial processes and poses serious risks to occupational safety and health. This study proposes an optimized machine-learning-based approach to predict CO concentration using metal-oxide semiconductor (MOS) sensor arrays. The model was trained and evaluated on a public dataset comprising 650 time-series measurements from 14 thermally modulated MOS sensors, tested across CO concentrations ranging from 0 to 8.9 ppm under dynamic relative humidity (15%–75%). To optimize computational efficiency and mitigate multicollinearity, a multi-method feature selection strategy that combines Random Forest importance, Recursive Feature Elimination (RFE), and Mutual Information (MI) was implemented, successfully isolating sensors R10, R11, and R13 as the most robust predictors. A Random Forest Regression model, optimized via grid search and validated through five-fold cross-validation, was subsequently developed. The proposed framework demonstrated high predictive accuracy, achieving an R^2 of 0.884, Root Mean Square Error (RMSE) of 2.189 ppm, Mean Absolute Error (MAE) of 1.215 ppm, and Symmetric Mean Absolute Percentage Error (SMAPE) of 34.27%. These results highlight the potential of combining low-cost, feature-optimized MOS sensor arrays with ensemble machine learning for accurate, real-time gas monitoring. The framework provides a computationally efficient decision-support tool for the early detection of hazardous CO levels, contributing to safer process environments.

INTRODUCTION

Carbon monoxide (CO) is a toxic, colorless, and odorless gas with numerous harmful effects. In chemical process industries, CO may emerge as a by-product of various reactions, such as methanol production and fossil fuel combustion [1]. High concentrations of CO pose significant safety risks not only to the working environment but also to the quality and safety of the resulting products [1]. In addition to its impact on occupational safety, prolonged exposure to low levels of CO has been shown to increase the risk of cardiovascular diseases, including heart attacks and strokes [2]. Furthermore, CO concentration is considered a potential biomarker for the early diagnosis of respiratory and neurodegenerative disorders [3]. Therefore, the development of CO detection technologies is essential to support the improvement of workplace safety, environmental monitoring, and public health.

Gas Chromatography–Mass Spectrometry (GC–MS) has been widely employed for CO detection due to its high sensitivity and accuracy [3]. However, GC-MS-based detection has limitations, including high cost, long processing time, and unsuitability for real-time field monitoring [3]. These challenges highlight the need for the development of real-time CO detection systems. Various types of gas sensors have been developed to detect and quantify gases in industrial, environmental, and biomedical applications. Commonly used sensor types include Conducting Polymer (CP), Metal Oxide Semiconductor (MOS), Mixed Potential Solid Electrolyte (MPSE), Catalytic Metal (CM), and Quartz Crystal Microbalance (QCM) sensors. Each sensor type offers unique advantages and limitations depending on the operational environment and target gas, as shown in Table 1 [3]. For instance, CP sensors are known for their high sensitivity and fast response time but are often vulnerable to environmental interferences and require special calibration [4]. MPSE sensors function across a broad range of temperature and humidity conditions but can be mechanically fragile. CM sensors provide excellent sensitivity but are demanding in terms of sample purity and handling. QCM sensors offer rapid response and high accuracy but are relatively expensive and sensitive to environmental vibrations [5] [6]. Among these, MOS sensors are

https://doi.org/10.25077/jnte.v15n1.1390.2026

widely adopted due to their low cost, high sensitivity, and fast response time. These sensors operate based on changes in electrical resistance upon exposure to target gas molecules, making them suitable for real-time detection. Although they require higher operating temperatures and have relatively high energy consumption, their ease of fabrication, miniaturization potential, and compatibility with digital systems make them an attractive choice for applications such as carbon monoxide (CO) detection in industrial settings. Therefore, this study adopts MOS sensors as the primary detection platform due to their practicality and established effectiveness in gas monitoring systems, especially CO detection

Table 1. Comparison of characteristics, advantages, and limitations of various gas sensor types

Ref	Sensor	Key Advantages	Limitations/Disadvantages
[3]	Metal Oxide Semiconductor (MOS)	Low cost; High sensitivity; Fast response; Easy fabrication	High operating temperature; High energy consumption
[4]	Conducting Polymer (CP)	High sensitivity; Fast response time	Vulnerable to environmental interferences; Requires special calibration
[3]	Mixed Potential Solid Electrolyte (MPSE)	Functions in broad temp/humidity ranges	Mechanically fragile
[3]	Catalytic Metal (CM)	Excellent sensitivity	Demanding sample purity and handling requirements
[5], [6]	Quartz Crystal Microbalance (QCM)	Rapid response; High accuracy	Relatively expensive; Sensitive to environmental vibrations

Recent advancements in machine learning (ML) have significantly enhanced the performance of gas detection systems. ML algorithms can model the non-linear, high-dimensional, and dynamic nature of sensor array responses. While previous studies have successfully applied techniques such as Support Vector Regression (SVR) and Artificial Neural Networks (ANNs) for gas classification and concentration estimation [7]. Most existing approaches utilize data from the entire, large-scale sensor array. This recent method often leads to high multicollinearity, increased computational overhead, and susceptibility to environmental noise, such as fluctuating humidity and temperature. Consequently, deploying these resource-heavy models for continuous, real-time process safety monitoring at the edge remains a significant research gap.

To address this gap, this study proposes an optimized machine learning-based approach to predict CO concentrations. The novelty of this work lies in the development of a multi-method feature selection framework that combines Random Forest Mean Decrease in Impurity (MDI), Recursive Feature Elimination (RFE), and Mutual Information (MI) to isolate the most informative and robust subset of thermally modulated MOS sensors under dynamic environmental conditions. By systematically eliminating redundant features, the proposed framework not only maintains high predictive accuracy but also significantly reduces computational complexity.

The model is designed to support process safety monitoring by enabling accurate, real-time estimation of CO levels in industrial environments. To ensure the validity and reproducibility of the study, a publicly available dataset developed by Javier Burgués et al. is employed [8], [9], [10]. This dataset contains time-series responses from 14 temperature-modulated MOS sensors exposed to dynamic mixtures of CO and humidity, along with environmental variables such as temperature, flow rate, and heater voltage. Its richness and real-world relevance provide a reliable foundation for training and evaluating the predictive model.

METHODS

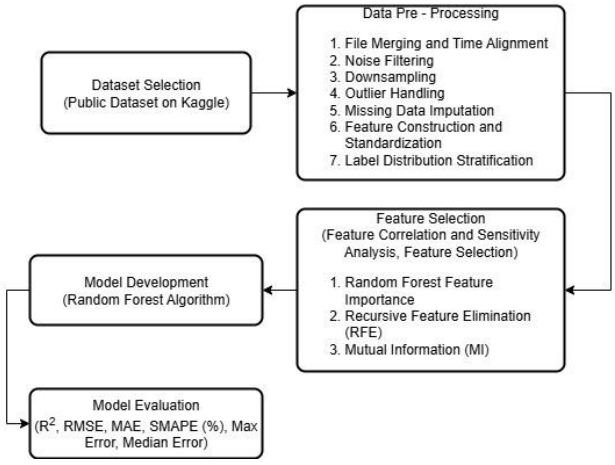


Figure 1. Research Method

Dataset Selection

This study utilized a publicly available dataset developed by Javier Burgués et al. [8], [9], [10], which consists of time-series measurements recorded from a gas sensing platform exposed to controlled mixtures of carbon monoxide (CO) and humid air. The dataset includes responses from 14 temperature-modulated metal oxide semiconductor (MOS) sensors: seven units of FIGARO TGS 3870-A04 and seven units of FIS SB-500-12. The sensors were installed on a custom-designed electronic board and exposed to a dynamic environment where CO and humidity levels were varied under a controlled gas flow system. A temperature and humidity sensor (SHT75, Sensirion) was used as a reference inside the test chamber. The heater voltage of the MOS sensors was modulated in cycles between 0.2–0.9 V, following a 25-second cycle as recommended by the manufacturer: 0.9 V for 5 s (high temperature), followed by 0.2 V for 20 s (low temperature), then repeating. To ensure response stability, sensors were pre-heated for one week before data acquisition.

Sensor resistance was recorded using voltage divider circuits with 1 MΩ load resistors, and the output voltage was sampled using an Agilent HP34970A/34901A DAQ at 15-bit precision and 3.5 Hz sampling rate. Temperature variations during each experiment were kept below 3°C, while relative humidity values were controlled in the range of 15% to 75%, chosen randomly for each sample. The experimental design aimed to simulate real-world operating conditions and to estimate the limit of detection (LOD) for CO. Based on recommendations from the IUPAC and literature on LOD calibration, CO concentrations were varied from 0 to 9 ppm, with five calibration points selected: [0.0, 2.2, 4.4, 6.7, 8.9 ppm]. Each concentration was tested in 10 repetitions, and each repetition had a randomly assigned humidity level drawn from a uniform distribution between 15% and 70%

https://doi.org/10.25077/jnte.v15n1.1390.2026

relative humidity. With 50 total conditions per day (5 concentrations × 10 repetitions), repeated over 13 separate days, the dataset comprises 650 measurements in total. This setup allows modeling and prediction of CO concentrations in a complex environment with realistic variability. The rich temporal resolution and environmental diversity in this dataset make it highly suitable for training and evaluating machine learning models for gas sensing applications.

Data Pre-Processing

All data manipulation, preprocessing, feature selection, and machine learning modeling in this study were executed using Python within a Jupyter Notebook environment. The Scikit-learn library was utilized for building the machine learning pipeline, while data handling and visualization were performed using Pandas and Matplotlib. Due to the high-frequency and multi-channel nature of the data, comprehensive preprocessing steps were applied to ensure signal quality and modeling reliability *[7]*.

File Merging and Time Alignment

The 13 daily measurement files were first merged into a single dataset. Time columns were standardized, and records were chronologically aligned to avoid overlaps or misordered samples caused by session-based acquisition resets.

Noise Filtering

To reduce high-frequency fluctuations caused by electrical noise and gas turbulence, a rolling average filter was applied to each sensor channel. A moving window of 10–20 samples (3–6 seconds) was selected empirically to preserve signal shape while smoothing transients *[11]*.

Downsampling

To simplify model training and reduce computational load, the data were downsampled to 1 Hz using a time-averaging method. This preserved the underlying signal dynamics while reducing data redundancy and memory usage *[3]*.

Outlier Handling

In gas sensor datasets, outliers can arise from a variety of sources such as electrical noise, hardware instability, calibration drift, or external environmental disturbances. These outliers may significantly distort statistical distributions and degrade the performance of machine learning models if not properly handled [12], [13].

To systematically identify outliers, we used the Interquartile Range (IQR) method, which is robust to non-Gaussian distributions and well-suited for noisy sensor data. For each feature (sensor resistance, temperature, humidity, or heater voltage), the first quartile (Q1) and third quartile (Q3) were calculated, and the IQR was defined in Equation 1.

$$IQR = Q3 - Q1 \tag{1}$$

Values lying outside the range as shown in Equation 2.

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR] \tag{2}$$

were considered outliers. These values were either:

- Removed if they were isolated and not part of a stable sequence, or
- Interpolated using linear or spline interpolation if surrounded by valid measurements within a short gap (e.g., <10 seconds), to preserve temporal continuity.

This approach helps maintain data integrity while minimizing the influence of anomalies on model training. It is widely recommended in sensor data literature for preprocessing time series with unstable baselines or low-frequency spikes [11], [13].

Missing Data Imputation

Missing data in the dataset was primarily caused by:

- Communication delays during high-speed data acquisition,
- Sensor warm-up phases,
- Occasional disconnections or saturation in environmental sensors.

To address this, short-duration missing values were interpolated linearly across time using surrounding data points. This technique assumes smooth signal transitions, which is generally valid for MOS sensor responses under slowly changing gas concentrations. For longer gaps (e.g., more than 60 consecutive seconds), data segments were excluded entirely to prevent the introduction of artificial trends or interpolation bias. This cutoff threshold was selected based on experimental cycle durations and the typical timescale of CO concentration changes. Preserving the temporal dynamics of sensor response is critical in gas sensing applications, especially under variable humidity and heating conditions. Linear interpolation remains a widely accepted method for imputation in gas sensor studies if gaps are short and the physical process remains smooth *[13], [14]*.

Feature Construction and Standardization

After preprocessing and cleaning, each data sample was structured into a feature vector composed of:

- **Sensor resistances (R1–R14)**: These are the raw resistance values (in MΩ) from the 14 MOS sensors.
- **Environmental parameters**: Temperature (°C), Relative Humidity (%), and Heater Voltage (V).

These features were used as input variables (independent variables) for the machine learning model, while CO concentration (ppm) served as the target output (dependent variable). To ensure that all features contribute equally to the learning process and are on a comparable scale, we applied z-score standardization to each feature. This transformation centers each feature around zero mean and scales it based on its standard deviation.

The z-score standardization is defined in Equation 3.

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

Where,

x is the original value of the feature,

μ is the mean of the feature values in the training set,

σ is the standard deviation of the feature in the training set,

z is the standardized value.

This transformation ensures that all features contribute equally to the model, preventing any single feature with a large numeric scale (such as resistance values in MΩ) from disproportionately influencing the learning process. It also helps improve training stability and accuracy for algorithms that are sensitive to feature scaling (e.g., Random Forest, Support Vector Regression, or neural networks) [7].

The standardization parameters (μ and σ) were calculated using only the training set to avoid data leakage. These same values were then used to transform the corresponding features in the test set, maintaining consistency across evaluation stages. This feature scaling step is commonly recommended in preprocessing pipelines for gas sensor data and other industrial sensor systems where multivariate measurements differ in magnitude and variability [5], [13].

Why standardization is important:

- Prevents features with larger numerical ranges (e.g., resistance in megaohms) from dominating the learning process.
- Essential for algorithms sensitive to feature magnitude (e.g., distance-based methods or regularized regression).
- Helps stabilize the optimization process and improve convergence in ensemble models like Random Forest and gradient boosting.

In this study, standardization was crucial to balance the influence of sensor outputs and environmental factors in predicting CO concentration under varying operational conditions.

Label Distribution Stratification

In supervised machine learning, especially regression tasks involving real-world experimental data, it is crucial to ensure that the target variable (label) is evenly distributed between the training and testing datasets. In this study, the target label was the CO concentration (ppm), which was experimentally varied across a predefined range of values. Although the experimental design aimed to test five CO concentration levels (0.0, 2.2, 4.4, 6.7, 8.9 ppm) uniformly with 10 repetitions each, real-world factors, such as transient delays, sensor saturation, or imperfect gas mixing, can introduce imbalanced or noisy label distributions over time. Without careful control, this can result in overrepresentation of some concentration levels in either the training or test set, potentially biasing the model and degrading its generalization performance. To address this, we employed label distribution stratification during the train-test splitting process. This technique ensures that the distribution of CO concentrations is preserved in both subsets, thereby maintaining representativeness across all measurement ranges. In our case, this was implemented using a stratified sampling strategy based on binned CO concentrations. More specifically:

- The continuous CO concentration values were first discretized into bins (e.g., via quantiles or fixed-width intervals),
- The dataset was then split using stratified sampling on these bins, preserving the relative frequency of each bin in both the training (80%) and testing (20%) datasets.

This approach is beneficial because:

- It minimizes sampling bias, particularly in small or imbalanced datasets,
- It ensures the model is exposed to the full range of CO levels during training,
- It enhances model robustness and fairness, especially for environmental sensing, where rare or extreme values (e.g., very low/high CO) are critical for safety.

Recent studies explained that stratification is especially valuable when modeling sensor systems with non-linear response profiles or when evaluating performance across different calibration points [2], [13].

Feature Selection

To identify the most relevant predictors of CO concentration, three feature selection techniques were applied, namely Random Forest Feature Importance, Recursive Feature Elimination, and Mutual Information.

Random Forest Feature Importance, which ranks features based on impurity reduction.

Random Forest (RF) is an ensemble learning algorithm that builds multiple decision trees during training and merges their outputs to improve predictive performance and reduce overfitting[15]. One of its key strengths is its ability to provide intrinsic feature importance measures, which quantify the contribution of each feature to the overall model performance.

Mean Decrease in Impurity (MDI)

This metric calculates how much each feature reduces the impurity (e.g., Gini index or variance) in a decision tree split. Features that consistently result in high information gain across trees are ranked higher. This method is computationally efficient and directly derived from the tree-building process. In recent research, MDI is especially useful in sensor-based systems where large multivariate inputs (e.g., 14 sensor channels) need to be ranked without requiring external models, so they used MDI to select the most relevant features for lithium-ion battery capacity estimation, successfully identifying the sensor signals most sensitive to capacity degradation [16].

Mean Decrease in Accuracy (MDA)

This permutation-based method evaluates the drop in model accuracy when each feature is randomly shuffled. A large decrease implies that the feature was important for prediction. While more computationally intensive, MDA is more robust to bias, especially in datasets with correlated features [15]. Although not implemented in this study, other research also shows that using ensemble-averaged feature importance across multiple RF models with dropout or bootstrapped subsamples improves the reliability of the ranking [17].

Recursive Feature Elimination (RFE), which iteratively removes the least important features using linear regression.

To identify the most relevant features for predicting CO concentration, this study employed Recursive Feature Elimination (RFE), a supervised wrapper-based technique that systematically removes irrelevant or redundant features by evaluating their contribution to model performance. RFE functions by training a regression model, specifically a Support Vector Regression (SVR) model with a linear kernel (C=1.0) used

https://doi.org/10.25077/jnte.v15n1.1390.2026

as the base estimator, on the entire set of input features. The elimination step size was set to 1, meaning the least important feature was removed iteratively during each cross-validation fold. The model evaluates the importance of each feature, and the least important feature is removed from the set. This process is repeated recursively until a predefined number of top-ranked features is retained. RFE is particularly useful in gas-sensing applications, where sensor signals often exhibit multicollinearity due to similar response characteristics under shared environmental conditions. However, conventional RFE may be biased when correlated features dominate the feature space, leading to the unintended exclusion of individually informative features. To address this limitation, advanced variants such as IRFS-SVR-RFE have been developed, as demonstrated by Xiong et al. (2024) [18], where features are first grouped based on correlation and then representative features are retained using a distance-based selection strategy. This approach enhances robustness and preserves critical but correlated information. Several studies across diverse domains have validated the effectiveness of RFE in improving model generalizability and reducing computational complexity, including the classification of cyberattacks in smart grid networks [19], partial charging analysis in lithium-ion batteries [20], hypothyroidism detection [21], and early dementia prediction using logistic regression enhanced with explainable artificial intelligence [22]. In the present study, RFE successfully reduced the original feature space of 17 variables by selecting a subset that preserved model accuracy while improving interpretability.

Mutual Information (MI), which captures nonlinear dependencies between features and the target.

Mutual Information (MI) is a powerful statistical measure used to evaluate the dependency between variables. In feature selection, MI quantifies the amount of information a given feature contributes to the target label or output variable. A high MI score between a feature and the label indicates that the feature is informative for prediction purposes. In traditional MI-based feature selection, features are ranked based on their individual mutual information values with respect to the class label. However, this approach may lead to the inclusion of redundant features if the inter-feature relationships are ignored. To overcome this, methods have been developed that consider both maximum relevance (high MI with the target) and minimum redundancy (low MI among selected features), often referred to as the mRMR (minimum Redundancy Maximum Relevance) criterion.

Recent advancements integrate neighborhood-based or fuzzy logic-based estimations of mutual information to better capture uncertainty and structure in high-dimensional datasets. For example, Sun et al. (2025) introduced an Adaptive Fuzzy Neighborhood Mutual Information (AFNMI) technique that enhances multilabel feature selection by adjusting neighborhood radii adaptively using Euclidean distances and constructing fuzzy neighborhood granules [23]. This method not only computes MI between features and labels but also incorporates label correlation and conditional mutual information to refine selection decisions. Similarly, Chen et al. (2025) developed a Normalized Mutual Information (NMI) strategy for driving state recognition, improving feature interpretability while accounting for the distributional properties of features [24]. Moreover, clustering-driven techniques such as Fuzzy C-means-based Weighted Neighborhood Mutual Information have been shown to enhance multi-label feature selection by evaluating feature relevance with respect to grouped label sets [23].

These methods demonstrate that incorporating adaptive, fuzzy, or normalized mutual information estimators can significantly improve the robustness and accuracy of feature selection strategies, especially in noisy, redundant, or multilabel data contexts. Features selected consistently across all methods were retained for final model training [13].

Machine Learning Model

The final regression model for predicting CO concentration was developed using the Random Forest Regression algorithm. Random Forest is a widely adopted ensemble learning technique that builds a collection of decision trees and aggregates their outputs to improve generalization, reduce variance, and handle nonlinear relationships in data. This method is particularly robust for modeling noisy sensor data and has demonstrated superior performance in high-dimensional environments where traditional models may suffer from overfitting or instability.

In this study, Random Forest was chosen due to its ability to rank feature importance, manage multicollinearity, and maintain performance even with partially redundant sensor signals. According to Heidari et al. (2023), Random Forest models are effective for feature selection and interpretation in complex environmental systems due to their built-in impurity-based ranking and ensemble averaging, and emphasize the utility of Random Forests combined with dropout and ensemble strategies to further stabilize predictions in the domain of fault detection and uncertainty modeling [17]. Moreover, as highlighted by Sin et al. (2025), Random Forest Regression has been successfully applied in scenarios requiring high robustness and generalization capability, including energy systems and battery health prediction [16].

The model in this study was trained on 80% of the total dataset and validated on the remaining 20% using a five-fold cross-validation strategy to minimize variance in performance estimation. To ensure rigorous optimization and reproducibility, hyperparameters were systematically tuned using a Grid Search method over a predefined parameter space. The search space encompassed: the number of trees (n_estimators) evaluated at [100, 300, 500, 700, 1000]; the maximum depth of the trees (max_depth) at [10, 15, 20, 25, 30, None]; and the maximum fraction of features to consider for splitting (max_features) at [0.5, 0.6, 0.7, 0.8, 1.0, 'sqrt']. The optimization objective was set to minimize the negative mean squared error (MSE) across the cross-validation folds [19]. This tuning process allowed the model to balance complexity and bias-variance trade-offs effectively. The final model performance was evaluated using standard regression metrics, and the feature importance scores were used to interpret the contribution of each sensor and environmental parameter to the prediction outcome.

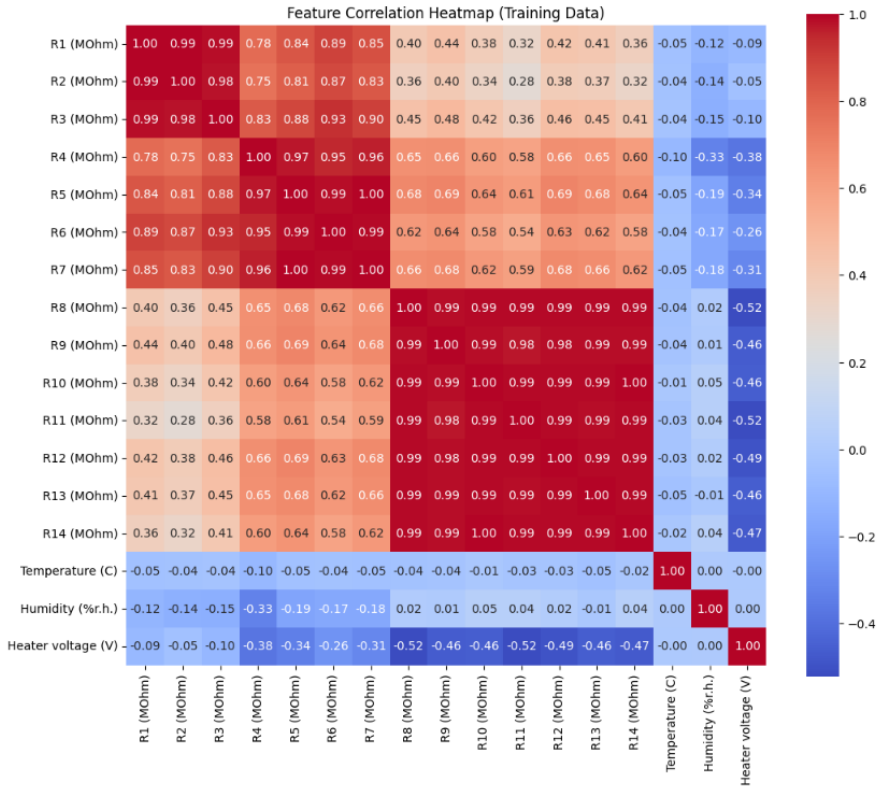


Figure 2. Correlation Analysis Using Heatmap

Model Evaluation

To evaluate the regression performance of machine learning models in estimating gas concentrations from sensor responses, the study employed six key statistical metrics:

Coefficient of Determination (R^2)

R^2 explains the proportion of variance in the dependent variable that is predictable from the independent variables. A value close to 1 indicates a strong fit.

$$R^2 = \frac{\sum_{i=1}^{N}(S_{act,i} - S_{pre,i})^2}{\sum_{i=1}^{N}(S_{act,i} - S_{act})^2} \quad (4)$$

Root Mean Square Error (RMSE)

RMSE quantifies the square root of the average squared differences between predicted and actual values. It penalizes larger errors more significantly than smaller ones, making it sensitive to outliers. A lower RMSE indicates higher predictive accuracy.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_{pre,i} - S_{act,i})^2} \quad (5)$$

Mean Absolute Error (MAE)

MAE measures the average magnitude of absolute errors between predicted and actual values. It provides a straightforward interpretation of model error in the same unit as the output.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|S_{pre,i} - S_{act,i}\right| \quad (6)$$

Symmetric Mean Absolute Percentage Error (SMAPE)

SMAPE expresses the prediction accuracy as a percentage by normalizing the absolute error using the average of actual and predicted values. This metric is especially useful when dealing with values near zero.

$$SMAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{\left|S_{pre,i} - S_{act,i}\right|}{(\left|S_{act,i}\right| + \left|S_{pre,i}\right|)/2} \quad (7)$$

https://doi.org/10.25077/jnte.v15n1.1390.2026

RESULTS AND DISCUSSION

Result

Feature Correlation

The correlation matrix analysis, as shown in Figure 2, revealed strong interdependencies among the MOS sensor features, particularly between sensors R8 and R14, which exhibited Pearson correlation coefficients exceeding 0.99. This high degree of multicollinearity is characteristic of sensor arrays using identical materials or operating conditions. While such consistency can ensure sensor robustness, it also introduces redundancy that can hinder model generalization and inflate variance during learning. Additionally, moderate negative correlations were observed between environmental variables (e.g., humidity and temperature) and sensor resistances, confirming the influence of ambient factors on sensor performance. This supports the inclusion of environmental parameters in the modeling pipeline.

Sensor Sensitivity Analysis

Sensor sensitivity analysis using linear regression coefficients, as depicted in Figure 3, highlights R10, R11, R13, and R7 as the most responsive sensors to variations in carbon monoxide (CO) concentrations. These sensors yielded the highest absolute coefficient magnitudes, indicating their dominant contribution to the predictive model. Such elevated weight values suggest that these specific MOS sensors are more effective in capturing the dynamic response to CO exposure, consistent with findings in prior studies involving thermally modulated MOS arrays [9].

In contrast, environmental variables such as heater voltage, temperature, and humidity, as well as sensors R12 and R14, exhibited comparatively low sensitivity. Although these features may not significantly influence the model independently, their inclusion remains essential to account for signal drift and ambient interferences. This aligns with best practices in multivariate sensor systems, where redundant or low-impact features still contribute to model stability and generalization.

Feature Selection

Table 2. Features Scores

Sensor	RF	RFE Rank	MI
R1 (Mohm)	0.072690	1	0.008411
R2 (Mohm)	0.052634	1	0.021211
R3 (Mohm)	0.014352	1	0.018391
R4 (Mohm)	0.015773	1	0.027249
R5 (Mohm)	0.016010	1	0.025362
R6 (Mohm)	0.009217	2	0.020275
R7 (Mohm)	0.010406	2	0.023415
R8 (Mohm)	0.023377	3	0.385337
R9 (Mohm)	0.027454	3	0.166267
R10 (Mohm)	0.116810	4	0.444709
R11 (Mohm)	0.434807	4	0.552016
R12 (Mohm)	0.022689	5	0.410680
R13 (Mohm)	0.057557	5	0.422859
R14 (Mohm)	0.015350	6	0.471055
Temperature (°C)	0.008842	6	0.166267
Humidity (%r.h.)	0.017821	7	0.628738
Heater Voltage (V)	0.084210	7	0.003790

Note: RF=Random Forest, RFE= Recursive Feature Elimination, MI=Mutual Information

To identify the most relevant predictors of carbon monoxide (CO) concentration, three feature selection techniques were applied: Random Forest Feature Importance (RF), Recursive Feature Elimination (RFE), and Mutual Information (MI). These techniques provide complementary perspectives: RF ranks features based on their contribution to impurity reduction in decision trees, RFE recursively removes the least informative features using linear regression, and MI estimates the nonlinear dependency between each input variable and the target variable. Table 2 summarizes the feature scores across the three methods. For instance, Heater Voltage (V) exhibited the highest importance score in RF (0.894210), while R13, R10, and R11 were also consistently ranked among the top features across different methods, indicating their strong association with CO concentration.

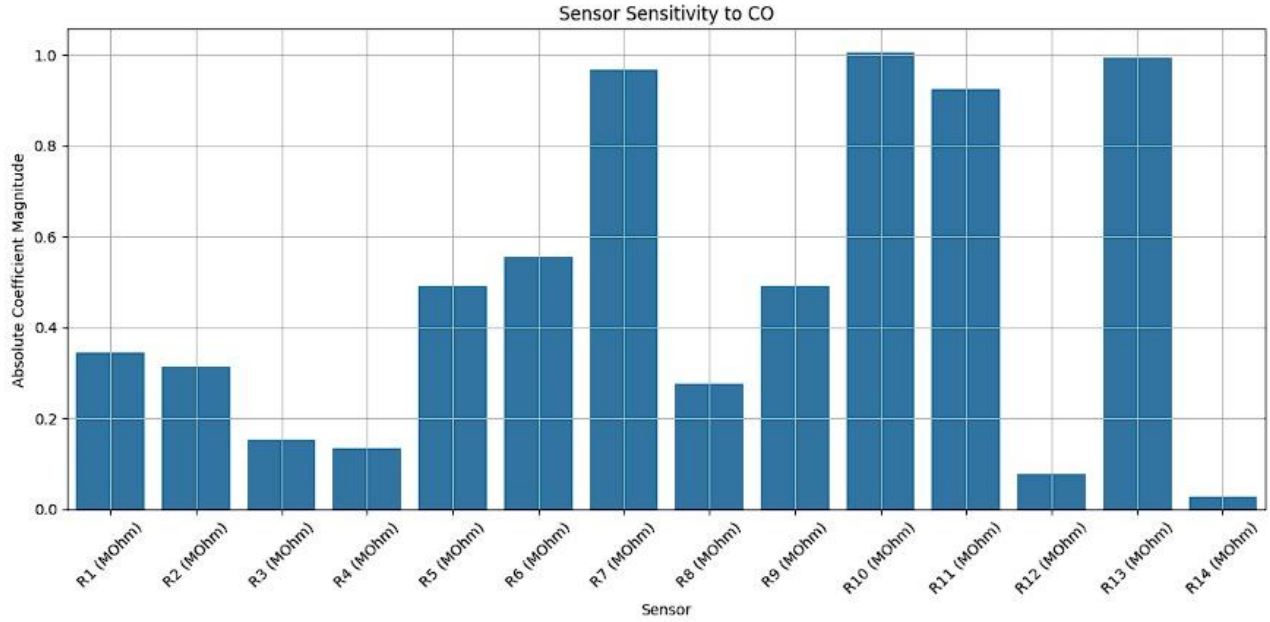


Figure 3. Sensor Sensitivity to CO

Table 3. Selected Features

Sensor	RF	RFE Rank	MI	Votes
R13 (Mohm)	True	True	True	3
R10 (Mohm)	True	True	True	3
R11 (Mohm)	True	True	True	3
R14 (Mohm)	-	-	True	1
R5 (Mohm)	-	True	-	1
Heater Voltage (V)	True	-	-	1
Humidity (%r.h.)	False	-	True	1
R7 (Mohm)	-	True	-	1
R1 (Mohm)	True	False	-	1
R12 (Mohm)	False	-	False	0
R2 (Mohm)	False	False	-	0
R8 (Mohm)	False	False	False	0
R6 (Mohm)	-	False	-	0
R4 (Mohm)	-	-	False	0
R9 (Mohm)	False	False	False	0
R3 (Mohm)	-	-	-	0
Temperature (°C)	-	-	False	0

Note: RF=Random Forest, RFE= Recursive Feature Elimination, MI=Mutual Information

To synthesize the results, a voting-based summary table was constructed, as shown in Table. In this table, each feature was labeled as True if it appeared in the top 5 most important features for a given method, if it ranked within the top 6–10 was labeled as False, and "-" if it was excluded from the top 10 entirely. The Votes column aggregates how many times a feature was marked "True" across the three methods. Features such as R13, R10, and R11 received a perfect score of 3 votes, indicating that they were selected in the top 5 by all methods. This consensus reinforces their relevance for inclusion in the final model. Conversely, features like R12, R2, and Temperature (°C) received zero votes, as they were not ranked among the top 10 by any method, suggesting limited or inconsistent predictive value. These features can be reasonably excluded from model training to reduce complexity without sacrificing performance. This multi-method feature selection approach enhances the robustness of the final predictor set by leveraging both linear and nonlinear relevance measures, thereby improving generalizability and reducing the risk of overfitting.

Model Development Using Random Forest

Following the hyperparameter tuning process described in the methodology, the optimal configuration for the Random Forest Regression model was identified. The parameters that yielded the best predictive performance during cross-validation are summarized in Table 4.

Table 4. Hyperparameters of The Model Development Using Random Forest

Hyperparameter	Value
max_depth	25
max_features	0.8
n_estimators	500
random_state	42

Model Evaluation

The performance of the Random Forest Regression model was rigorously evaluated using multiple quantitative metrics, each offering a different lens into the model's predictive behavior. As shown in Table 5, the error matrix reflects the predictive performance of the trained Random Forest Regression model for estimating CO concentration. The coefficient of determination (R^2) is 0.884, indicating that approximately 88.4% of the variance in the target variable (CO concentration) can be explained by the selected input features. This demonstrates strong goodness-of-fit and suggests that the model effectively captures the underlying relationships in the data.

Table 5. Model Evaluation

Error Matrix	Value
R^2	0.884
RMSE	2.189
MAE	1.215
SMAPE(%)	34.271%

The Root Mean Square Error (RMSE) is 2.189, which quantifies the average magnitude of the residuals between the predicted and actual values. Lower RMSE values indicate better model performance, and in this context, the RMSE suggests moderate prediction error. The Mean Absolute Error (MAE) is 1.215, representing the average of the absolute differences between predicted and actual values. Compared to RMSE, MAE is less sensitive to outliers, and this value suggests the model provides reliable point estimates across most predictions. The Symmetric Mean Absolute Percentage Error (SMAPE), reported at 34.271%, measures relative prediction accuracy by comparing the absolute error to the average of the predicted and actual values. While SMAPE is particularly useful for percentage-based performance assessment, a value above 30% may indicate room for improvement in model calibration or feature selection, particularly under varying environmental conditions.

Overall, these evaluation metrics indicate a reasonably accurate and stable model, though further refinements, such as noise filtering or advanced feature engineering, may help reduce error margins and improve generalization.

Discussion

The present study demonstrates the potential of metal oxide semiconductor (MOS) sensors combined with machine learning to predict carbon monoxide (CO) concentrations as a means to support process safety monitoring. The feature selection results, obtained using Random Forest Importance, Recursive Feature Elimination (RFE), and Mutual Information (MI), consistently identified sensors R13, R10, and R11 as the most influential predictors. This convergence across three distinct selection methods enhances confidence in the relevance of these sensors and justifies their inclusion in the final model [9]. The physical mechanism underlying this high sensitivity is rooted in the surface chemistry of the MOS sensing layer. When exposed to CO, a reducing gas, the molecules react with chemisorbed oxygen ions on the heated MOS surface, releasing trapped electrons back into the semiconductor's conduction band and causing a measurable drop in electrical resistance. The consistent selection of R10, R11, and R13 indicates that these sensors have optimal sensing-layer compositions or are positioned in temperature-modulated phases that maximize the catalytic reaction.

Further investigation using correlation analysis revealed a high degree of collinearity among resistance-based sensor signals, particularly among R1 through R10. Despite this multicollinearity, the use of ensemble methods such as Random Forest mitigates the risk of overfitting due to their inherent

https://doi.org/10.25077/jnte.v15n1.1390.2026

robustness to redundant features [7]. Furthermore, the moderate negative correlation observed with environmental features is physically justified; water vapor (humidity) competes with oxygen for adsorption sites on the MOS surface, which typically dampens sensor sensitivity. By capturing these non-linear interactions, the model effectively compensates for environmental drift.

The sensitivity analysis corroborated the importance of R7, R10, R11, and R13, which displayed the largest regression coefficients, aligning with the earlier feature ranking results. The model, trained with a Random Forest Regressor, achieved an R^2 score of 0.884, with an RMSE of 2.189 and MAE of 1.215. To provide an analytical comparison, this predictive performance demonstrates a significant advantage over traditional linear calibration models, which frequently struggle to map the non-linear drift caused by humidity and temperature variations in MOS arrays. While baseline linear regression applied to raw sensor data often yields lower accuracy due to uncompensated cross-sensitivity and multicollinearity, the proposed Random Forest algorithm mitigates this by utilizing its ensemble structure to capture complex, multi-dimensional sensor interactions. Furthermore, compared to resource-intensive deep learning models, this RFE-optimized Random Forest achieves highly competitive accuracy (R^2 = 0.884) while maintaining a lightweight computational architecture. This makes it highly viable for real-time edge computing and integration into existing Internet of Things (IoT) safety monitoring systems in process plants.

However, the SMAPE value of 34.271% suggests a relatively higher variation in percentage error, which may be attributed to low-range CO concentrations or residual sensor drift effects that were not fully compensated during preprocessing [25]. Despite promising results, this study has several limitations. The use of public datasets restricts control over experimental variability, and the model was only validated under laboratory conditions. Additionally, environmental disturbances such as sensor aging, cross-sensitivity to other gases, or long-term drift were not addressed explicitly, potentially limiting the real-world applicability of the findings.

CONCLUSIONS

This study successfully presents an optimized data-driven framework for predicting carbon monoxide (CO) concentrations to enhance process safety monitoring using metal oxide semiconductor (MOS) gas sensors. By applying a robust multi-method feature selection strategy that combines Random Forest importance ranking, Recursive Feature Elimination (RFE), and Mutual Information (MI), the original high-dimensional sensor array was systematically reduced to the three most informative predictors: sensors R10, R11, and R13. This quantifiable reduction in the feature space minimized computational complexity and mitigated multicollinearity without sacrificing predictive performance. Using this optimized subset, the developed Random Forest Regression model achieved a high coefficient of determination ($R^2 = 0.884$) alongside low residual errors ($RMSE = 2.189$, $MAE = 1.215$), demonstrating its capability to reliably estimate CO concentrations under variable environmental conditions. These measurable achievements underscore the viability of combining targeted subsets of MOS sensors with ensemble machine learning to create cost-effective, computationally lightweight gas-monitoring systems. For future research directions, it is essential to validate this optimized model through real-time testing in actual industrial process environments. Subsequent work will focus on integrating this predictive machine learning framework into Internet of Things (IoT) monitoring architectures for continuous, edge-computing applications. Additionally, addressing long-term sensor degradation (drift) and extending the methodology to multi-gas detection scenarios will be critical next steps toward developing comprehensive, real-world early hazard detection systems for industrial safety.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to colleagues and fellow researchers for their valuable support, constructive feedback, and insightful discussions throughout the development of this study. Their encouragement and critical perspectives significantly contributed to the improvement of the research quality and clarity of this work.

REFERENCES

[1] K. Bhardwaj *et al.*, "Carbon monoxide detection, separation, and conversion into valuable products via smart nanomaterials: Challenges and perspectives," *Materials Today*, vol. 83, pp. 404–445, Mar. 2025, doi: 10.1016/j.mattod.2024.12.020.

[2] L. Xue, J. Dang, X. Li, F. Liu, Z. Qin, and Q. Wang, "Trace detection of carbon monoxide by AACVD of Ni-doped ZnO nanosheets," *Sensors and Actuators B: Chemical*, vol. 442, p. 138126, Nov. 2025, doi: 10.1016/j.snb.2025.138126.

[3] Z. Zhang *et al.*, "Electronic nose based on metal oxide semiconductor sensors for medical diagnosis," *Progress in Natural Science: Materials International*, vol. 34, no. 1, pp. 74–88, Feb. 2024, doi: 10.1016/j.pnsc.2024.01.018.

[4] F. Raspagliesi, G. Bogani, S. Benedetti, S. Grassi, S. Ferla, and S. Buratti, "Detection of Ovarian Cancer through Exhaled Breath by Electronic Nose: A Prospective Study," *Cancers*, vol. 12, no. 9, p. 2408, Aug. 2020, doi: 10.3390/cancers12092408.

[5] K. Liu and C. Zhang, "Volatile organic compounds gas sensor based on quartz crystal microbalance for fruit freshness detection: A review," *Food Chemistry*, vol. 334, p. 127615, Jan. 2021, doi: 10.1016/j.foodchem.2020.127615.

[6] N. M. Zetola *et al.*, "Diagnosis of pulmonary tuberculosis and assessment of treatment response through analyses of volatile compound patterns in exhaled breath samples," *Journal of Infection*, vol. 74, no. 4, pp. 367–376, Apr. 2017, doi: 10.1016/j.jinf.2016.12.006.

[7] Pradyumn, P. B. Barman, A. Sil, and S. K. Hazra, "Recent advancement in selective gas sensors and role of machine learning," *Journal of Alloys and Compounds*, vol. 1030, p. 180757, May 2025, doi: 10.1016/j.jallcom.2025.180757.

[8] J. Burgués, J. M. Jiménez-Soto, and S. Marco, "Estimation of the limit of detection in semiconductor gas sensors through linearized calibration models," *Analytica Chimica Acta*, vol. 1013, pp. 13–25, Jul. 2018, doi: 10.1016/j.aca.2018.01.062.

[9] J. Burgués and S. Marco, "Multivariate estimation of the limit of detection by orthogonal partial least squares in temperature-modulated MOX sensors," *Analytica*

Chimica Acta, vol. 1019, pp. 49–64, Aug. 2018, doi: 10.1016/j.aca.2018.03.005.
[10] L. Fernandez, J. Yan, J. Fonollosa, J. Burgués, A. Gutierrez, and S. Marco, "A Practical Method to Estimate the Resolving Power of a Chemical Sensor Array: Application to Feature Selection," *Front. Chem.*, vol. 6, p. 209, Jun. 2018, doi: 10.3389/fchem.2018.00209.
[11] M. Al-Hashem, S. Akbar, and P. Morris, "Role of Oxygen Vacancies in Nanostructured Metal-Oxide Gas Sensors: A Review," *Sensors and Actuators B: Chemical*, vol. 301, p. 126845, Dec. 2019, doi: 10.1016/j.snb.2019.126845.
[12] S. T. Araya *et al.*, "Performance assessment of machine learning techniques in electronic nose systems for power transformer fault detection," *Energy and AI*, vol. 20, p. 100497, May 2025, doi: 10.1016/j.egyai.2025.100497.
[13] J. Xie, M. Sage, and Y. F. Zhao, "Feature selection and feature learning in machine learning applications for gas turbines: A review," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105591, Jan. 2023, doi: 10.1016/j.engappai.2022.105591.
[14] A. S. AlSalehy and M. Bailey, "Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset," *Smart Cities*, vol. 8, no. 3, Art. no. 3, Jun. 2025, doi: 10.3390/smartcities8030082.
[15] M. G. L. Brown, M. G. Peterson, I. K. Tezaur, K. J. .Peterson, and D. L. Bull, "Random forest regression feature importance for climate impact pathway detection," *Journal of Computational and Applied Mathematics*, vol. 464, p. 116479, Aug. 2025, doi: 10.1016/j.cam.2024.116479.
[16] S. Sin, E. Kang, J. Kim, S. Oh, and J. Baek, "Random forest-based small training dataset complementation and features selection for capacity estimation of lithium-ion batteries in electric-powered application," *Applied Soft Computing*, vol. 181, p. 113526, Sep. 2025, doi: 10.1016/j.asoc.2025.113526.
[17] M. Heidari, M. H. Moattar, and H. Ghaffari, "Forward propagation dropout in deep neural networks using Jensen–Shannon and random forest feature importance ranking," *Neural Networks*, vol. 165, pp. 238–247, Aug. 2023, doi: 10.1016/j.neunet.2023.05.044.
[18] J. Xiong, T. Ma, Z. Lian, and R. de Dear, "Perceptual and physiological responses of elderly subjects to moderate temperatures," *Building and Environment*, vol. 156, pp. 117–122, Jun. 2019, doi: 10.1016/j.buildenv.2019.04.012.
[19] O. Kornyo *et al.*, "Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms," *Computers & Security*, vol. 135, p. 103456, Dec. 2023, doi: 10.1016/j.cose.2023.103456.
[20] F. Tian, S. Chen, X. Ji, J. Xu, M. Yang, and R. Xiong, "Robust lithium-ion battery state of health estimation based on recursive feature elimination-deep Bidirectional long short-term memory model using partial charging data," *International Journal of Electrochemical Science*, vol. 20, no. 1, p. 100891, Jan. 2025, doi: 10.1016/j.ijoes.2024.100891.
[21] N. S. Santhosh, K. M, K. K, and A. S, "AGXLG: Enhancing Hypothyroidism Prediction with Recursive Feature Elimination and Boosting Techniques," *Procedia Computer Science*, vol. 258, pp. 3457–3467, 2025, doi: 10.1016/j.procs.2025.04.602.
[22] R. Ahmed *et al.*, "A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction," *Healthcare Analytics*, vol. 6, p. 100362, Dec. 2024, doi: 10.1016/j.health.2024.100362.
[23] L. Sun, J. Guo, X. Wu, and J. Xu, "Fuzzy C-means clustering-based multi-label feature selection via weighted neighborhood mutual information," *Information Sciences*, vol. 718, p. 122389, Nov. 2025, doi: 10.1016/j.ins.2025.122389.
[24] J. Chen, F. Fan, C. Wei, K. Polat, and F. Alenezi, "Decoding driving states based on normalized mutual information features and hyperparameter self-optimized Gaussian kernel-based radial basis function extreme learning machine," *Chaos, Solitons & Fractals*, vol. 199, p. 116751, Oct. 2025, doi: 10.1016/j.chaos.2025.116751.
[25] N. N. Viet, P. H. Phuoc, L. V. Thong, N. V. Chien, and N. Van Hieu, "A comparative study of machine learning models for identifying noxious gases through thermal fingerprint measurements and MOS sensors," *Sensors and Actuators A: Physical*, vol. 375, p. 115510, Sep. 2024, doi: 10.1016/j.sna.2024.115510.

NOMENCLATURE

Meaning of symbols used in the equations and other symbols presented in your article must be presented in this section.

CO	meaning of	Carbon Monoxide
MOS	meaning of	Metal Oxide Semiconductor
MI	meaning of	Mutual Information
RF	meaning of	Random Forest
RFE	meaning of	Recursive Feature Elimination
RMSE	meaning of	Root Mean Square Error
MAE	meaning of	Mean Absolute Error
SMAPE	meaning of	Symmetric Mean Absolute Percentage Error
R^2	meaning of	Coefficient of Determination
V	meaning of	Voltage
℃	meaning of	Degrees Celsius

AUTHOR(S) BIOGRAPHY

Hermin Kartika Sari, S.T., M.Eng.
Hermin Kartika Sari, S.T., M.Eng. is a lecturer in the Department of Chemical Engineering at Politeknik Negeri Bandung. She earned both her bachelor's and master's degrees in Physics Engineering from Universitas Gadjah Mada. Currently, she is pursuing a Doctoral program in Electrical Engineering at Universitas Gadjah Mada. Her research interests encompass instrumentation, industrial instrumentation, and environmental instrumentation. Committed to advancing applied science and engineering, Hermin actively engages in teaching, research, and the development of technology-based solutions for environmental and industrial challenges. She can be reached via email at hermin.kartika@polban.ac.id.

Thomas Oka Pratama, S.T., M.Eng.
Thomas Oka Pratama, S.T., M.Eng., is a distinguished lecturer in the Department of Physics Engineering at Universitas Gadjah Mada. He holds a bachelor's and master's degree in physics engineering from Universitas Gadjah Mada. Currently, he is pursuing a Doctoral program in Environmental Science at Universitas Gadjah Mada. His research interests primarily focus on environmental instrumentation, sustainable technology, and sensor systems for environmental monitoring. With a deep commitment to fostering innovation in sustainable engineering, Thomas actively contributes to both academic teaching and

https://doi.org/10.25077/jnte.v15n1.1390.2026

research development. His work is dedicated to solving real-world environmental issues through integrated technological solutions. He can be reached via email at thomas.o.p@ugm.ac.id.

Yohana Fransiska Ferawati, S.T., M.T.
Yohana Fransiska Ferawati, S.T., M.T., is a specialist in the field of Materials. With a strong expertise in Materials in Chemical Engineering, Yohana has a keen research interest in the development and analysis of novel material properties for various applications. Her contributions to both academia and research are highly valued. She can be reached via email yohana.fransiska@polban.ac.id

Gita Nur Sajida, S.T., M.T.
Gita Nur Sajida, S.T., M.T., is a dedicated researcher and lecturer in the field of Catalysis. With a chemical engineering background, Gita focuses on the development and application of catalysts for efficient chemical reactions and industrial processes. Her passion for innovation and academic excellence is reflected in her significant contributions to teaching and research. She can be reached via email gita.nur.sajida@polban.ac.id

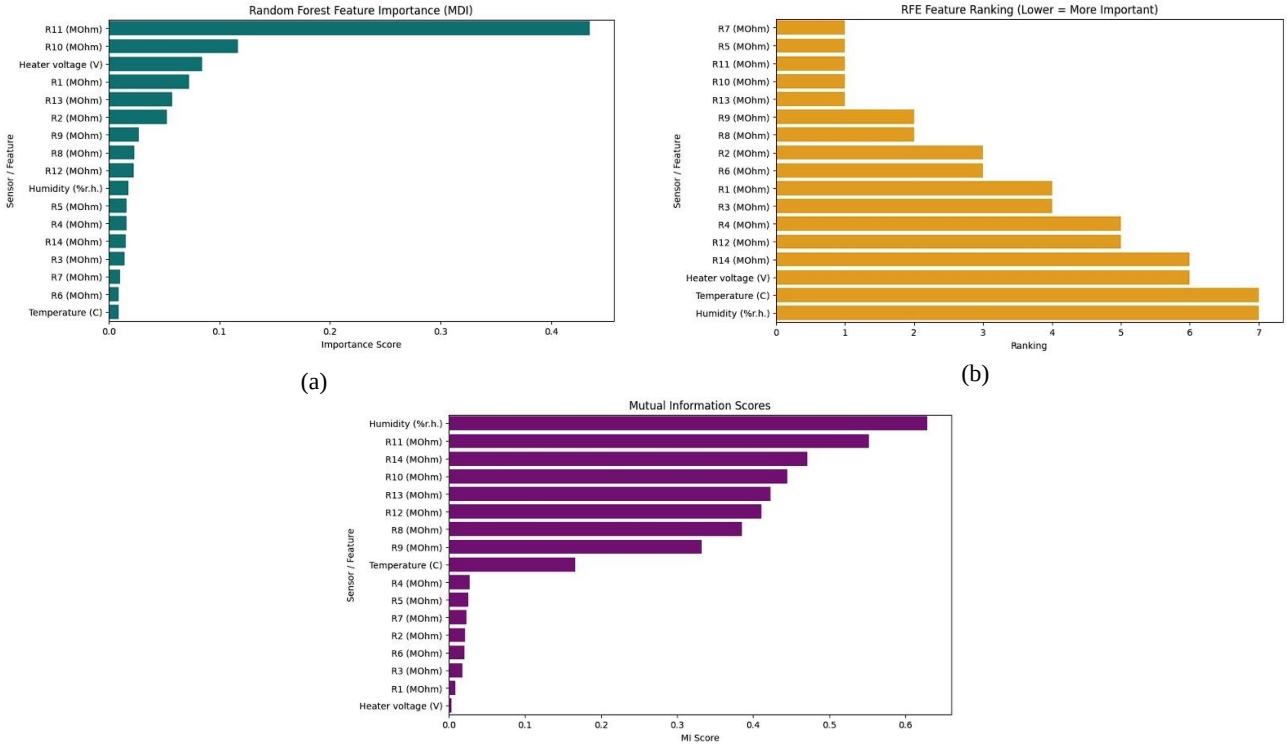
Gustin Mustika Krista, S.S.T., M.T.
Gustin Mustika Krista, S.S.T., M.T., is an expert in the field of Bioprocesses. With a background in Chemical Engineering, Gustin focuses on the development and optimization of biological processes. His dedication to innovation in bioprocesses has made significant contributions to both research and industrial applications. He can be reached via email gustin.mustika@polban.ac.id

Teguh Taufiqurohim, S.T., M.T.
Teguh Taufiqurohim, S.T., M.T., is an expert in Energy and Renewable Technologies. Focusing on sustainable energy solutions, Teguh has a strong interest in the research and development of innovative technologies for the future of energy. His contributions to advancing the science and application of renewable energy are highly significant. He can be reached via email teguh.taufiqurohim@polban.ac.id

Dr. Shoerya Shoelarta, LRSC, M.T.
Dr. Shoerya Shoelarta, LRSC, M.T. is a lecturer in the Department of Chemical Engineering at Politeknik Negeri Bandung. He holds a master's degree in chemical engineering and a doctorate with research focused on catalysis and process intensification. His areas of expertise include green synthesis, catalytic processes, microwave-assisted reactions, renewable energy conversion, and process simulation. Dr. Shoelarta has actively contributed to scientific advancements in biodiesel production, zeolite synthesis, graphene material development, and energy efficiency auditing in geothermal power plants. His scholarly work reflects a commitment to sustainable chemical engineering practices and applied research in industrial processing. He can be contacted via email at shoerya.shoelarta@polban.ac.id

APPENDICES

Figure 4 shows the bar chart comparison of sensor feature importance using three selection methods, namely Mutual Information, Random Forest Importance, and Recursive Feature Elimination, highlighting the most relevant predictors of CO concentration.



(a) (b)

(c)

Figure 4. The Bar Chart Comparison of sensor feature importance using three selection methods: (a) Rndom Forest Feature Importance, (b) Recursive Feature Elimination, and (c) Mutual Information

https://doi.org/10.25077/jnte.v15n1.1390.2026