Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 63-83
DOI: 10.38016/jista.1566965

Outliers Treatment for Improved Prediction of CO and NOx Emissions from Gas Turbines Using Ensemble Regressor Approaches

Vahid Sinap[1*]

[1] Department of Management Information Systems, Ufuk University, Ankara, Türkiye

vahidsinap@gmail.com

Abstract
Gas turbines are widely used in power generation plants due to their high efficiency, but they also emit pollutants such as CO and NOx. This study focuses on developing predictive models for predicting CO and NOx emissions from gas turbines using machine learning algorithms. The dataset used includes pollutant emission data from a combined cycle gas turbine (CCGT) in Türkiye, collected hourly between 2011 and 2015. Various outlier treatment methods such as Z-Score, Interquartile Range (IQR), and Mahalanobis Distance (MD) are applied to the dataset. Machine learning algorithms including Random Forest, Extra Trees, Linear Regression, Support Vector Regression, Decision Tree, and K-Nearest Neighbors are used to build the predictive models, and their performances are compared. Additionally, Voting Ensemble Regressor (VR) and Stacking Ensemble Regressor (SR) methods are employed, using Gradient Boosting, LightGBM, and CatBoost as base learners and XGBoost as a meta-learner. The results demonstrate that the SR model, when applied to the dataset processed using the IQR method, achieves the highest prediction accuracy for both NOx and CO emissions, with R^2 values of 0.9194 and 0.8556, and RMSE values of 2.7669 and 0.4619, respectively. These findings highlight the significant role of the IQR method in enhancing model accuracy by effectively handling outliers and reducing data noise. The improved data quality achieved through this method contributes to the superior performance of the SR model, making it a reliable approach for predicting NOx and CO emissions with high precision.

Keywords: Gas turbine emissions, Machine learning, Outlier processing, Combined cycle power generation, Interquartile range, Mahalanobis distance

Gaz Türbinlerinden Kaynaklanan CO ve NOx Emisyonlarının Tahmininde Aykırı Değer İşleme ve Topluluk Regresyon Yaklaşımlarının Kullanımı

Öz
Gaz türbinleri, yüksek verimlilikleri nedeniyle enerji üretim tesislerinde yaygın olarak kullanılmaktadır; ancak, aynı zamanda CO ve NOx gibi zararlı gaz emisyonlarına da neden olmaktadırlar. Bu çalışma, gaz türbinlerinden kaynaklanan CO ve NOx emisyonlarını tahmin etmek için makine öğrenmesi algoritmalarını kullanarak tahmin modelleri geliştirmeye odaklanmaktadır. Kullanılan veri seti, Türkiye'deki bir kombine çevrim gaz türbininden (CCGT) 2011 ve 2015 yılları arasında saatlik olarak toplanan emisyon verilerini içermektedir. Veri setine Z-Skoru, Çeyrekler Arası Aralık (IQR) ve Mahalanobis Mesafesi (MD) gibi çeşitli aykırı değer işleme yöntemleri uygulanarak modellerin performansına etkisine incelenmiştir. Modeller oluşturulurken Rastgele Orman, Ekstra Ağaçlar, Doğrusal Regresyon, Destek Vektör Regresyonu, Karar Ağacı ve K-En Yakın Komşu gibi makine öğrenmesi algoritmaları kullanılmış ve performansları karşılaştırılmıştır. Ayrıca, Gradient Boosting, LightGBM ve CatBoost algoritmalarını temel temel öğrenici ve XGBoost'u meta-öğrenici olarak kullanan Oylama Topluluk Regresyonu (VR) ve İstifleme Topluluk Regresyonu (SR) yöntemlerinin de performansları incelenmiştir. Sonuçlar, IQR yöntemiyle işlenen veri seti üzerinde uygulanan SR modelinin hem NOx hem de CO emisyonları için en yüksek tahmin doğruluğunu sağladığını göstermektedir. Modelin R^2 değeri NOx için 0.9194, CO için 0.8556 olarak bulunmuş; RMSE ise sırasıyla 2.7669 ve 0.4619 olarak elde edilmiştir. IQR yöntemiyle elde edilen iyileştirilmiş veri kalitesi, SR modelinin üstün performans göstermesine katkı sağlamakta ve modelin NOx ve CO emisyonlarını yüksek hassasiyetle tahmin edebilmesi açısından güvenilir bir yaklaşım olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Gaz türbini emisyonları, Makine öğrenmesi, Aykırı değer işleme, Kombine çevrim enerji üretimi, Çeyrekler arası aralık, Mahalanobis uzaklığı

* Corresponding Author.
E-mail: vahidsinap@gmail.com

Recieved : 14 Oct 2024
Revision : 24 Jan 2025
Accepted : 5 Feb 2025

1. Introduction

Gas turbines (GT) are a widely preferred energy conversion technology in power generation plants due to their high efficiency and reliability. Simply put, GT consist of a series of turbines and compressors on a shaft rotating at high speed. GT take in air from outside and compress it using a compressor. The compressed air is then mixed with a fuel (usually natural gas or oil) and sent to a combustion chamber. In the combustion chamber, the fuel-air mixture is ignited and burns under high temperature and pressure. The high-pressure and high-temperature gases generated by this combustion expand and pass through the turbine, causing the turbine blades to rotate. This rotational movement turns the turbine shaft, which is connected to a generator, thus generating electrical energy. The open cycle of a GT is illustrated in Figure 1.

While the high efficiency and reliability of GT make them an ideal choice for electricity generation, harmful gases such as carbon monoxide (CO) and nitrogen oxides (NOx) released during the combustion process cause environmental impacts. CO is produced as an incomplete combustion product when there is not enough oxygen, or the combustion process is incomplete. CO is a colorless, odorless gas and can be dangerous to humans. When inhaled at dangerous levels, CO can cause severe poisoning and even death (Liu et al. 2021). NOx is the general term for compounds formed because of the reaction of nitrogen and oxygen in the atmosphere under high temperature and pressure, including nitrogen monoxide (NO) and nitrogen dioxide (NO_2). NOx can contribute to acid rain, ozone formation, and air pollution, causing respiratory diseases and environmental damage (Pandey and Chandrashekhar, 2014).

Various methods and technologies are used to reduce the release of harmful gases such as CO and NOx from GT into the environment. These include exhaust gas treatment systems such as selective catalytic reduction (SCR) and selective non-catalytic reduction (SNCR). The SCR system lowers NOx emissions by injecting ammonia (NH_3) or urea into the exhaust gas stream. In the presence of a catalyst, these substances react with NOx, converting it into harmless nitrogen (N_2) and water (H_2O) (Wardana and Lim, 2022). The SNCR system reduces NOx by injecting ammonia or urea into the exhaust gas without the use of a catalyst, relying on high temperatures to facilitate the reaction (Mahmoudi et al. 2010). Additionally, improving combustion efficiency and optimizing the air-fuel ratio can significantly reduce the formation of CO and NOx emissions (Tian et al. 2024). For example, low NOx combustion techniques can be used to provide higher combustion efficiency while minimizing NOx formation. Techniques such as water or steam injection can be used to lower combustion temperatures, thereby reducing NOx emissions. In addition, regular maintenance and cleaning of GT can help reduce CO and NOx emissions. Other effective strategies for reducing CO and NOx emissions include utilizing cleaner fuels, enhancing combustion chamber design, and implementing exhaust gas recirculation (EGR) systems (Kumar et al. 2022). However, traditional emission control methods often face disadvantages such as high costs, complexity, and efficiency issues. Implementing and operating large-scale exhaust gas treatment systems often involve significant costs (Lott et al. 2024). Additionally, the environmental impact of some technologies must be considered. For example, certain exhaust gas treatment systems can produce harmful by-products that may be released into the environment (Lopes et al. 2015). This necessitates a broad and comprehensive evaluation of emission control processes on an ongoing basis to ensure environmental sustainability and regulatory compliance.

In recent years, machine learning (ML) has become a prominent technology for evaluating emission control processes by predicting emissions from GT. ML algorithms analyze large amounts of data and take into account various variables such as operating conditions of GT, fuel composition, air temperature and other environmental factors. By identifying complex relationships between these variables, they create models to predict emissions from GT. These models can quickly respond to changes in the operating conditions of GT and keep emission predictions up to date. For instance, if there is a sudden change in the operating conditions of a gas turbine and its impact on emissions is immediately identified and assessed, control measures can be taken automatically if necessary. Furthermore, predictive models allow proactive measures to be taken, considering operating conditions. ML models can predict future operating conditions by analyzing historical performance data and environmental conditions of the plant. In a scenario where the developed model predicts an impending air temperature increase and identifies its potential impacts on CO and NOx emissions, power plant operators can adjust plant operating parameters based on this information or take proactive measures to minimize emissions by making a specific process change. In this way, predictive models can guide power plants to develop and implement strategies to reduce environmental impacts.

The promising advantages of ML in predicting emissions from GT have been evaluated by some important studies in the literature. Aslan (2024) evaluates the performance of machine learning models, including AdaBoost, XGBoost, and Random Forest (RF), in predicting gas turbine emissions. Using random search optimization, the study finds that AdaBoost achieves the highest accuracy (99.97%) and lowest mean square error (MSE = 2.17). Dirik (2022) conducted a study using the Adaptive Neural Fuzzy Inference System (ANFIS) method to model and predict NOx emissions from a natural gas-fired combined cycle power plant (CCPP). The results demonstrated that the ANFIS models achieved high accuracy in predicting

NOx emissions. Pachauri (2024) discusses the importance of monitoring harmful gas emissions from GT in CCPPs, particularly CO and NOx, to ensure compliance with emission standards. The study proposes a stacked ensemble machine learning (SEM) model for predicting CO and NOx emissions from a CCPP gas turbine. The model uses neural network for regression (NNR), generalized additive model (GAM), and bagging of regression trees (BT) as base learners, with a generalized regression neural network (GRNN) as a meta-learner. The hyperparameters of SEM are optimized using a Bayesian optimization algorithm. The performance of SEM is compared with support vector regression (SVR), decision tree (DT), and linear regression (LR). Simulation results show that SEM significantly reduces the root mean square error (RMSE) for NOx and CO compared to other ML techniques, demonstrating its higher predictive accuracy. The study by Kochueva and Nikolskii (2021) investigates the utility of predictive emission monitoring systems (PEMS) as software solutions to validate and complement costly continuous emission monitoring systems for natural gas electrical generation turbines. The research focuses on building a model for predicting CO and NOx emissions based on ambient variables and technological process parameters using various ML methods. The developed models achieve coefficients of determination of $R^2 = 0.83$ for NOx emissions and $R^2 = 0.89$ for CO emissions. In their study, Kaya et al. (2019) introduce a novel PEMS dataset collected over a period of five years from a gas turbine, specifically for the predictive modeling of CO and NOx emissions. The data is analyzed using a contemporary ML approach, providing valuable insights into emission predictions. It is noted in the study that the most successful algorithm model for the exhaust gas emission prediction is Extreme Learning Machines (ELM). In the study conducted by Dalal et al. (2023), commonly used ML regression models such as Multiple Linear Regression (MLR), DT, RF, Adaboost Regressor, Gradient Boosting Regressor (GB), and XGBoost Regressor were compared using the same dataset for predicting emissions like CO and NOx. According to the results of the research, the RF Model showed the best performance with the highest accuracy of 0.60 for NOx prediction and 0.65 for CO prediction. Coelho et al. (2024) conducted a study to estimate CO and NOx emissions from a gas turbine using the PEMS dataset. They employed four feature generation methods: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE). Various regression models, including Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), K-Nearest Neighbors (KNN), Cubist Regression, RF, Light Gradient Boosting Machine (LGBM), Categorical Boosting, and Deep Forest Regression (DFR), were evaluated with all the generated features. The DFR model achieved the best results, with an R^2 value of 0.53 for CO emissions in the validation dataset. For NOx emissions, the DFR model achieved an R^2 value of 0.47 for the validation dataset. The study by Yousif et al. (2024) aims to predict gas emissions from natural gas power plants. A hybrid model combining Feed Forward Neural Network (FFNN) and Particle Swarm Optimization (PSO) was developed for this purpose. The FFNN predicts NOx and CO emissions, while the PSO optimizes the FFNN weights to enhance prediction accuracy. The PSO employs a unique random number selection strategy using the KNN algorithm. Neighbor Component Analysis (NCA) is used to select parameters most correlated with emissions. The model was tested with publicly available datasets and evaluated using MSE, mean absolute error (MAE), and RMSE metrics. Results show that the PSO significantly improves FFNN training, increasing CO and NOx prediction accuracy by 99.18% and 82.11%, respectively. Naghibi (2024) develops an advanced gas turbine forecasting model using ensemble decision trees and robust preprocessing. The bagging structure outperforms boosted trees, achieving a lower RMSE (1.4176) with fewer estimators. While effective overall, the model has limitations in specific operating ranges. The study offers insights for optimizing gas turbine efficiency and improving electricity supply reliability.

The aim of this study is to develop predictive models for predicting CO and NOx emissions from GT using algorithms such as RF, Extra Trees (ET), LR, SVR, DT, and KNN, and to compare their performances with VR and SR methods. In VR and SR methods, GB, LightGBM and CatBoost algorithms are used as base learners and XGBoost algorithm is used as meta learner. The models are trained on a dataset containing emission data collected over a five-year period (01/01/2011–31/12/2015). Unlike previous studies that used the same dataset, this research focuses on the problem of outliers in the dataset. By examining and comparing the effects of various outlier treatment methods such as Z-Score, IQR, and MD on the developed models, this study aims to provide a novel contribution to literature. Additionally, the findings of this study are expected to contribute to the field of emission prediction from GT and provide valuable insights for environmental management in the energy sector. By enhancing the accuracy of the developed models, the study aims to play a critical role in environmental impact assessments and sustainable energy policies.

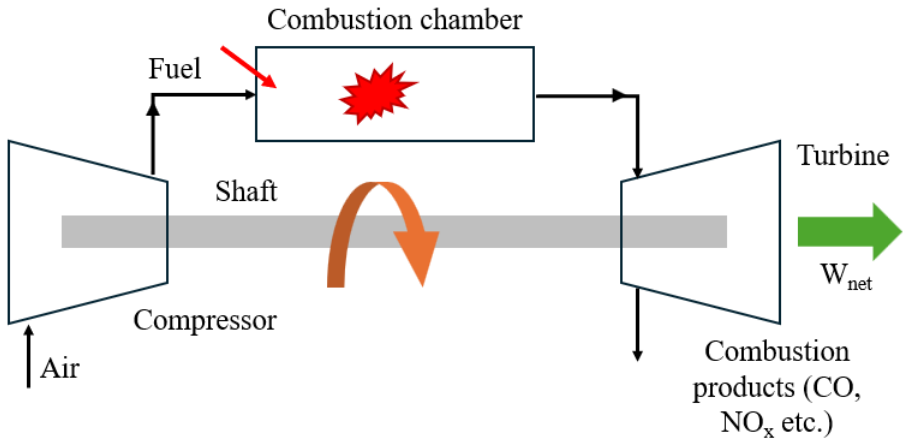

Figure 1. Open cycle of a GT

2. Methodology

In this section, the characteristics of the dataset used in the study, descriptions of the ML algorithms, performance criteria used in the comparison of the algorithms and information about the data preparation process are given. With this information, it is aimed to establish the methodological and analytical foundations of the research, to increase the scientific contribution of the study and to ensure its reproducibility.

2.1. Dataset

The dataset used in this study includes pollutant emission data from a CCGT in Türkiye. The dataset consists of sensor data collected hourly between 2011 and 2015 and is openly available through the UCI repository (Kaya et al. 2019). This set, which includes 36,733 data records in total, belongs to the periods when the power plant operated between 75% and 100% load factors. A graphical representation of the output features is given in Figure 2.

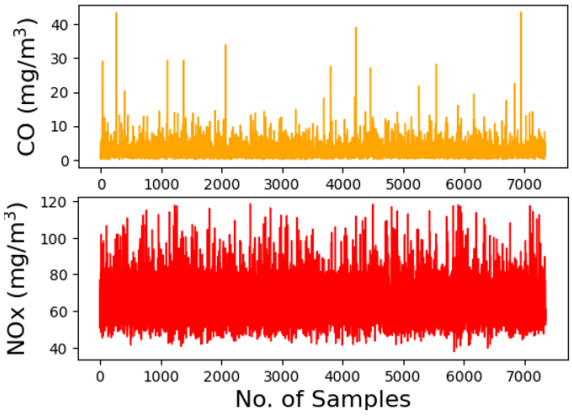

Figure 2. Visual depiction of the output features NOx and CO.

The dataset includes various environmental and operational parameters that affect the performance of the gas turbine. These parameters include ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), air flow differential pressure (AFDP), gas turbine exhaust pressure (GTEP), compressor discharge pressure (CDP), turbine energy yield (TEY), and turbine inlet temperature (TIT) and turbine afterburner temperature (TAT). The main pollutants produced by GT are CO and NOx, while sulfur oxides (SOx) and other pollutants vary depending on the type of fuel used. Table 1 presents the statistical analysis of two output variables (CO and NOx) and nine input variables (AT, AP, AH, AFDP, GTEP, TIT, TAT, CDP and TEY) in the dataset. Especially the atmospheric parameters such as AT, AP, and AH play an important role in predicting CO and NOx emissions (Farzaneh-Gord, and Deymi-Dashtebayaz, 2011). For CCGT, AFDP, GTEP, TIT, TAT, and CDP parameters are very influential, and sensor locations and measurement methods of these variables are of great importance (Wood, 2023). AFDP sensors are usually placed before and after the air filter, GTEP sensors in the exhaust duct, TIT sensors at the turbine inlet, TAT sensors at the turbine outlet, and CDP sensors at the compressor outlet. Correct positioning and regular calibration of these sensors ensures efficient and safe operation of CCGT systems.

Understanding the relationships between input and output variables is critical for improving the accuracy of predictive models. These relationships are analyzed using the correlation coefficient (CC), which is calculated with Pearson correlation. CC values indicate the level of dependence between variables, with positive values indicating a direct relationship and negative values indicating an inverse relationship. Figure 3 shows the correlations between CO, NOx, and other input variables. The concentration of CO demonstrates negative correlations with several operational parameters, including AT, AFDP, GTEP, TIT, TEY, and

CDP, with correlation coefficient values of -0.17, -0.45, -0.52, -0.71, 0.57, and -0.55, respectively. This implies that as the turbine's inlet temperature and the compressor's discharge pressure decrease, the emission of CO increases. Conversely, NOx emissions exhibit a higher level of correlation with a decrease in AT (-0.56). During the winter season, it is recommended to operate the gas turbine at higher temperatures to mitigate NOx emissions. However, GT operation is also negatively correlated with AFDP (-0.19), GTEP (-0.20), TIT (-0.21), TAT (-0.09), TEY (-0.12), and CDP (-0.17), respectively. Moreover, CO demonstrates positive correlations with AP (0.07), AH (0.11), and TAT (0.06), while NOx is positively correlated with AP and AH, with correlation coefficient values of 0.19 and 0.16, respectively. Figure 4 illustrates the schematic diagram of the CCPP, encompassing all input and output features.

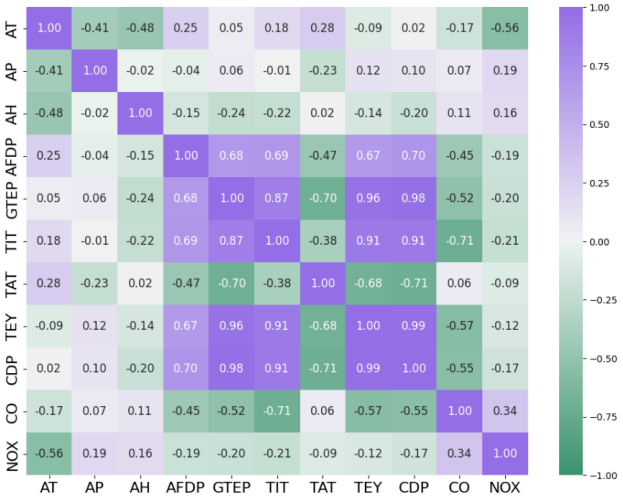

Figure 3. Matrix of Pearson correlation coefficients among the features

Table 1. Statistical overview of the dataset

Features	Unit	Average	Min	Max	Skewness	Kurtosis	Standard Dev.
AT	°C	17.71	-6.23	37.10	-0.0435	-0.8266	7.4474
AP	mbar	1013.07	985.85	1036.60	0.1941	0.4419	6.4633
AH	%	77.86	24.08	100.20	-0.6280	-0.2745	14.4613
AFDP	mbar	3.92	2.08	7.61	0.3810	0.2246	0.7739
GTEP	mbar	25.56	17.69	40.71	0.3290	-0.6538	4.1959
TIT	°C	1081.42	1000.80	1100.90	-0.8882	-0.0457	17.5363
TAT	°C	546.15	511.04	550.61	-1.7559	2.0167	6.8423
CDP	mbar	12.06	9.85	15.15	0.2367	-0.6315	1.0887
TEY	MWh	133.50	100.02	179.50	0.1165	-0.5001	15.6186
CO	mg/m^3	2.37	0.0003	44.10	4.8381	49.0817	2.2626
NOx	mg/m^3	65.29	25.90	119.91	1.0267	2.0375	11.6783

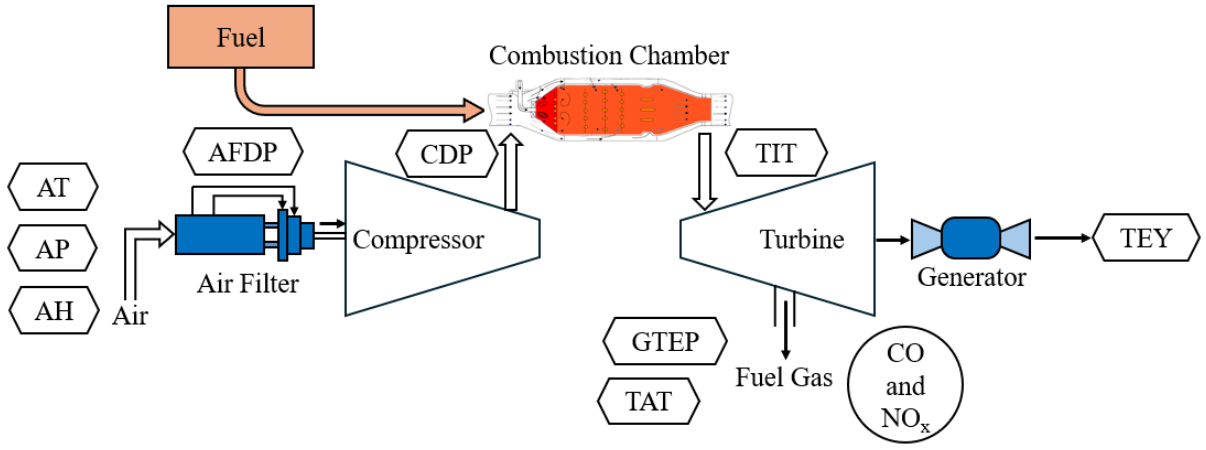

Figure 4. The schematic diagram illustrating the CCPP includes all input and output features

2.2. Data preparation

In any ML or data analysis project, data preparation is a critical step that significantly influences the accuracy and performance of the resulting models. Proper data preparation involves cleaning, transforming, and organizing the raw data into a format suitable for analysis. This process helps to ensure that the models are trained on high-quality data, which is essential for achieving reliable and meaningful predictions. It also involves handling missing values, removing outliers, and normalizing data, all of which contribute to the robustness of the analysis. In this study, careful attention has been paid to the data preparation phase to maximize the predictive performance of the models for CO and NOx emissions from a gas turbine. The dataset was carefully inspected, and no null or missing values were

found, ensuring the dataset remained complete and representative. To test the data's compliance with the normal distribution assumption, the Shapiro-Wilk test was applied.

Figure 5 presents box plots for the various input and output features used in this study, providing a visual summary of their distributions. The box plots illustrate the central tendency and variability of each variable, as well as the presence of any potential outliers. For instance, variables like AP and AH exhibit a relatively tight IQR, indicating low variability, whereas variables such as CO and NOx show a wider IQR, signifying higher variability and the presence of numerous outliers.

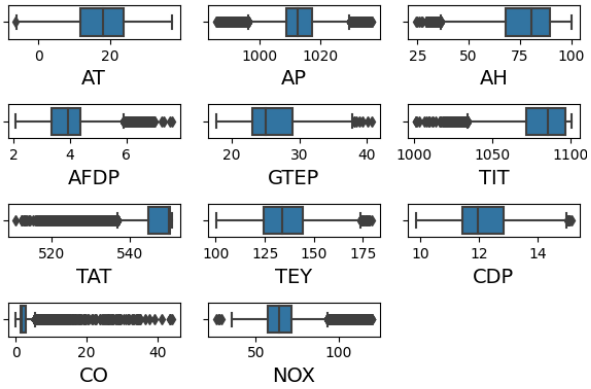

Figure 5. Box plots of input and output features showing distribution and potential outliers

The study focuses on the impact of different outlier treatment methods on the performance of the developed forecasting models. The outlier handling methods examined include Z-Score, IQR and MD. Each method offers a unique approach to identifying and handling outliers, which can significantly impact model accuracy and reliability.

- **Z-Score**

The Z-Score method identifies outliers based on the number of standard deviations a data point is from the mean. The Z-Score for a data point x is calculated using the formula in Equation 1.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the mean of the data and σ is the standard deviation. Data points with Z-Scores greater than a specified threshold (commonly ± 3) are considered outliers. This method assumes that the data follows a normal distribution, and it is particularly useful for detecting extreme values in symmetric distributions. However, the Z-Score method is sensitive to the assumption of normality. If the data is not normally distributed, the Z-Score method may incorrectly identify outliers. Additionally, it is less effective for small datasets or datasets with high variability, as the mean and standard deviation can be heavily influenced by extreme values.

- **Interquartile Range**

IQR method identifies outliers based on the spread of the middle 50% of the data. The IQR is calculated as the difference between the third quartile $(Q3)$ and the first quartile $(Q1)$ given in Equation 2.

$$IQR = Q3 - Q1 \tag{2}$$

Outliers are typically defined as data points that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$. This method is robust to non-normal distributions and is effective in handling skewed data. The IQR method is less sensitive to extreme values compared to the Z-Score method, but it may not be as effective for datasets with a small number of observations, as the quartiles may not accurately represent the data distribution. The choice of the multiplier (e.g., 1.5 or 3) can also affect the number of outliers detected, requiring careful tuning.

- **Mahalanobis Distance**

The MD method identifies outliers by considering the distance of a data point from the mean of the distribution, taking into account the correlations between variables. The MD for a data point x is given as in Equation 3.

$$D^2 = (x - \mu)^{T\,\Sigma^{-1}} \quad (x - \mu) \tag{3}$$

where μ is the mean vector of the data, and Σ is the covariance matrix. Data points with a MD exceeding a certain threshold (determined by the chi-square distribution with degrees of freedom equal to the number of variables) are considered outliers. This method is particularly effective for multivariate data and can identify outliers that may not be evident when considering variables individually. However, the MD method is sensitive to the distribution of the dataset. If the dataset is small or homogeneous, the MD method may incorrectly identify outliers, leading to overfitting (Caicedo et al. 2017). The method assumes that the data follows a multivariate normal distribution, and if this assumption is violated, the MD may not accurately identify outliers. In datasets with high dimensionality, the MD method can be computationally expensive and may struggle with the "curse of dimensionality," where the distance metric becomes less meaningful. For small and homogeneous datasets, the MD method may overfit the model by identifying extreme data points as outliers, which can lead to a model that performs well on the training data but poorly on new or unseen data. In

datasets with high variability or noise, the MD method may misinterpret the variance and flag some data points as outliers, which can weaken the generalization ability of the model.

2.3. Regression algorithms

In this study, RF, ET, LR, SVR, DT and KNN algorithms were used to predict CO and NOx emissions from GT. RF is an ensemble learning algorithm consisting of many decision trees. Each decision tree is trained on subsets of randomly selected features and data samples. This increases the generalization ability of the model and makes it more resistant to overfitting. Final predictions are usually made by averaging the predictions of these trees. The basic idea of RF is based on the idea that many different and random trees can come together to form a more powerful model (Biau, 2012). In this way, the errors within each tree compensate for each other and a better prediction can be made overall. The formula for RF for regression is given in Equation 4.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \tag{4}$$

Here, $\hat{y}$ is the predicted value, N is the total number of decision trees and $h_i(x)$ is the prediction of the i-th decision tree. With this formula, the final prediction is calculated by averaging the predictions of all trees.

ET is an ensemble learning algorithm similar to RF, but with certain differences. ET aims to increase diversity by generating decision trees in a more randomized way. When building decision trees, the best split point for each node is randomly selected. This allows the trees to be more diverse from each other, which helps the ensemble model to become more generalizable. It has been observed that ET can have faster training times compared to RF (Ahmad et al. 2018).

LR is a basic regression algorithm used to model the relationship between dependent and independent variables. This algorithm attempts to capture the linear relationship between the values of the independent variables in the dataset and the dependent variable (Maulud and Abdulazeez, 2020). The model determines the coefficients of the features in the dataset and a constant (cut-off point). These coefficients and constant represent the linear relationship that will best explain the observations in the dataset. LR is particularly effective when the dependent variable is continuous and there is a linear relationship between the variables. The LR formula is presented in Equation 5.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{5}$$

When Equation 5 is analyzed, $\hat{y}$ represents the predicted value, β_0 represents the cut-off point, β_j represents the coefficient of the j-th independent variable and x_j represents the value of the j-th independent variable.

SVR is an adaptation of the Support Vector Machines (SVM) algorithm for regression analysis. While SVM was originally developed for classification problems, SVR modifies it to address regression tasks. SVR employs the concept of a hyperplane used in SVM for classifying data points, but in this case, the hyperplane is determined to ensure that the data points lie within a specified margin (Valkenborg et al. 2023). The basic idea of SVR is to fit the data points, i.e., the training data, around a hyperplane in such a way that the hyperplane is positioned to provide the widest margin possible. This margin is defined as the distance between the hyperplane and the closest data points on either side. However, unlike classification where data points are expected to be separated by the hyperplane, in regression, it is unrealistic to expect all data points to lie exactly on the hyperplane. Instead, SVR aims to find a balance between fitting the data points closely while maintaining a margin of tolerance, allowing for some deviation from the hyperplane within a defined threshold (Yu and Kim, 2012). Therefore, a tolerance (ε) margin is defined and the hyperplane tries to classify data points within this margin. Other data points may fall outside the tolerance margin. SVR determines the regression line by minimizing a cost function, which can also be controlled by hyperparameters "C" and "ε". The parameter "C" controls the model's resistance to overfitting, while the parameter "ε" determines the tolerance margin. The regression formula for SVR is presented in Equation 6.

$$\underset{\mathbf{w},\, b}{min}\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*) \tag{6}$$

Here, $\mathbf{w}$ represents the weight vector, b the bias term, C the regularization parameter and $\varepsilon_i,\ \varepsilon_i^*$ the error terms. This formulation determines the hyperplane according to the maximum margin principle while minimizing the error terms for data points that lie within the specified tolerance margin.

DT is a regression analysis technique that uses features and outcomes from the dataset. It works based on decision trees and each tree is used to predict outcomes based on the values of features in the data set. The algorithm starts by creating a decision node for each feature in the dataset. These nodes test the feature values in the dataset and create smaller subsets by partitioning the data according to a specific rule. Each node tries to choose the best feature and threshold value to best partition the dataset. The tree-building process continues until the dataset is divided by a certain criterion (for example, until a certain depth or minimum number of samples is reached). As a result, each leaf node makes a

prediction, and the average or weighted average of these predictions is used as the tree's prediction based on the values of the features in the dataset (Quinlan, 1996). DT provides a flexible modeling method to capture complex relationships. In Equation 7, the basic estimation formula for DT is given.

$$\hat{y} = \frac{1}{N_t} \sum_{x_i \in R_t} y_i \quad (7)$$

In this formula, the predicted value $\hat{y}$ is calculated as the average of the actual values of the samples in a leaf node. During this calculation, the number of samples in the leaf node N_t, the leaf node R_t and the actual value y_i for each sample are used.

KNN is a fundamental classification and regression algorithm. In classification, the class of a data point is determined by the majority vote of its k nearest neighbors. In regression, the output value of a data point is predicted by taking the average of the values of its k nearest neighbors. KNN relies on the similarity between data points and is generally considered a simple and effective method. However, it can be sensitive to noise and high-dimensional data issues and is often computationally expensive because it requires comparing the target data point against the entire training dataset to make predictions (Song et al. 2017). Equation 8 defines the regression formula for the KNN algorithm.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \quad (8)$$

In the formula, the estimate $\hat{y}$ is calculated as the average of the true values y_i of the k nearest neighbors. In this calculation, k neighbors (k) and the true value for each neighbor (y_i) are used.

2.4. Ensemble learning

In this study, the ensemble learning methods VR and SR are used. In these methods, GB, LightGBM and CatBoost algorithms are used as base learners, while XGBoost is chosen as the meta-learner to combine their predictions. While selecting the base learners and the meta-learner, the diversity, performance, and compatibility of the base learners were taken into account. Also, the overall impact of the meta-learner on the ensemble was considered. Ensemble methods aim to improve the overall performance of the model by combining the predictions of more than one base learner.

The VR combines the predictions of different base learners to generate the final prediction. The number of base learners is set to N and y^{ij} denotes the prediction of base learner i about sample j. In this case, the final prediction $y^{ensemble}$ of the VR is calculated as in Equation 9.

$$y^{ensemble} = \frac{1}{N} \sum_{i=1}^{N} y^{ij} \quad (9)$$

In the formula in Equation 9, $y^{ensemble}$ is defined as the final prediction and y^{ij} is defined as the prediction of the base learner i about the sample j.

A SR creates the final prediction by training a meta-learner on the predictions of base learners. The meta-learner is trained on a new dataset formed by the predictions of the base learners, and this meta-learner then takes the base learners' predictions as inputs to produce the final prediction (Divina et al. 2018). For example, if there are three different base learners, the predictions from these base learners form a new dataset. This dataset consists of the predictions of each base learner for each sample. Then, the meta-learner is trained on this new dataset along with the true values. The meta-learner uses the predictions of the base learners and the true values to make a more accurate prediction. The final prediction of the SR is the prediction made by the meta-learner. This method enhances the performance of the ensemble model by transforming the base learners' predictions into a structure that can model more complex relationships. The steps involved in the SR process are as follows:

1. Base learners' predictions: The predictions of the base learners are denoted as y^{ij}, where i represents the i-th base learner and j represents the index of the j-th sample.
2. Creating a new dataset: A new dataset is created using the predictions of the base learners. This new dataset is used to train the meta-learner. For each sample, the new dataset contains the predictions of all the base learners. Thus, the new representation of a sample's predictions can be denoted as $X^j = [y_{1j}, y_{2j}, \ldots, y_{nj}]$.
3. Meta-learner prediction: The meta-learner makes predictions using this new dataset and the true values. The prediction of the meta-learner is denoted as $\hat{y}_{\text{meta}}$.
4. Final prediction: The final prediction of the SR is achieved using the meta-learner's prediction, which is $\hat{y}_{\text{meta}}$.

Once this process is formalized, the final estimate of the SR is calculated as in Equation 10.

$$y^{ensemble} = \hat{y}_{\text{meta}} \quad (10)$$

2.5. Validation method

Many ML models rely on splitting datasets into training and testing to measure their performance. However, this method can lose reliability in terms of

accuracy as the size of the dataset used to test a small portion of the dataset decreases. In this study, the Stratified K-Fold Cross-Validation (SKCV) method is used to evaluate the generalization ability of the model. In the SKCV method, the dataset is divided into k equal parts. Each part is respectively selected as the test set, while the remaining k-1 parts are used as the training set. This process is repeated k times, and the model is trained and tested in each iteration. The overall performance of the model is evaluated by averaging the obtained performance metrics. This method can be more robust to noisy datasets and scatter of data points, which can better reflect the performance of the model on real-world data (Prusty et al. 2022). The basic formula of the SKCV method is given in Equation 11.

$$\text{SKCV} = \frac{1}{k}\sum_{i=1}^{k} L\left(y_i, \hat{y}_{-i}\right) \tag{11}$$

In the formula, k represents the number of layers ($k = 10$ in this study), y_i represents the true values in each layer, and $\hat{y}_{-i}$ represents the predicted values of the model trained outside that layer. The function L is used to measure the error between the true and predicted values. Usually, the mean squared error is used for regression problems or zero-one loss for classification problems.

2.6. Performance metrics

Evaluating research and validating its results requires the use of specific measurements and metrics. These metrics are used to evaluate the success of the model or algorithm and to understand its performance. This study examines performance metrics commonly used in regression problems such as R^2, RMSE and MSE. R^2 is a measure of a model's ability to explain variance in observed values. Its formula is given in Equation 12.

$$\text{R}^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{12}$$

In the formula in Equation 12, SS_{res} represents the residual sum of squares and SS_{tot} represents the total sum of squares. SS_{res} is defined by the formula in Equation 13 and SS_{tot} is defined by the formula in Equation 14.

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{14}$$

In the formulas, y_i represents the actual values, $\hat{y}_i$ the predicted values and $\bar{y}$ the mean of the observed values. R^2 takes values between 0 and 1 and the higher it is, the better the model explains the observed data.

RMSE is a metric that measures how far the model's predictions are from the true values. By taking the square root of the prediction errors, it shows the magnitude of the errors on average. Its formula is expressed in Equation 15.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{15}$$

In the RMSE formula, y_i represents the actual values, $\hat{y}_i$ represents the values predicted by the model, and n represents the total number of data points. The steps to calculate the RMSE are as follows:

1. For each observation, the difference (error) between the predicted value and the actual value is calculated: $e_i = y_i - \hat{y}_i$
2. The squares of these errors are taken: ${e_i}^2 = (y_i - \hat{y}_i)^2$
3. The mean of all the squared errors is computed: $\frac{1}{n}\sum_{i=1}^{n} {e_i}^2$
4. The square root of this mean is taken to find the RMSE.

MSE is the square of RMSE and represents the mean squared error of the model. The formula for MSE is given in Equation 16.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{16}$$

In the formula, y_i is the actual values, $\hat{y}_i$ is the values predicted by the model, and n is the total number of data points.

2.7. Model setups

For the development of the models, the data set was divided into two parts, 80% training and 20% testing. The randomness factor was set to 42 in all algorithms. The best hyperparameter settings for the models were determined using Bayesian Optimization. This method aims to discover the optimal parameter combinations using knowledge in the search domain. Bayesian Optimization is optimized to improve the performance of the model, working particularly effectively in complex and high-dimensional hyperparameter search spaces (Yang and Shami, 2020). This method offers a more efficient alternative to the classical Grid Search and Random Search techniques.

Table 2 lists the best hyperparameter settings determined using Bayesian Optimization for the ML models used in the study. The RF model was configured with 100 estimators, a maximum depth of 10, and

minimum samples of 2 to split a node and 1 to be at a leaf node, balancing complexity and generalization to prevent overfitting. The ET model, which randomizes tree generation, used 150 estimators, a maximum depth of 12, and minimum samples of 4 to split and 2 at a leaf, enhancing diversity and capturing complex patterns. For LR, fit_intercept was set to True to calculate the intercept, while normalize was set to False, as the data was preprocessed, ensuring the model captures linear relationships without unnecessary normalization. The SVR model used C = 1.0, ε = 0.1, and an 'rbf' kernel to handle non-linear relationships, with C controlling complexity and ε defining the error tolerance margin. The DT model was set with a maximum depth of 12 and minimum samples of 4 to split and 3 at a leaf, controlling tree growth to avoid overfitting. The KNN model used 10 neighbors, a 'distance' weighting function, and the 'ball_tree' algorithm to efficiently handle high-dimensional data, prioritizing closer neighbors for predictions. The VR combined Gradient Boosting, LightGBM, and CatBoost with equal weights, leveraging multiple algorithms for robust predictions. Finally, the SR used the same base learners as VR, with XGBoost as the meta-learner, creating a hierarchical model that captures complex relationships and achieves higher predictive accuracy by combining the strengths of multiple algorithms.

Table 2. Hyperparameter settings for ML models

Model	Hyperparameter	Settings
RF	n_estimators, max_depth, min_samples_split, min_samples_leaf	100, 10, 2, 1
ET	n_estimators, max_depth, min_samples_split, min_samples_leaf	150, 12, 4, 2
LR	fit_intercept, normalize	True, False
SVR	C, epsilon, kernel	1.0, 0.1, 'rbf'
DT	max_depth, min_samples_split, min_samples_leaf	12, 4, 3
KNN	n_neighbors, weights, algorithm	10, 'distance', 'ball_tree'
VR	estimators, weights	[('gb', GradientBoostingRegressor()), ('lgbm', LGBMRegressor()), ('cat', CatBoostRegressor())], [1, 1, 1]
SR	estimators, final_estimator	[('gb', GradientBoostingRegressor()), ('lgbm', LGBMRegressor()), ('cat', CatBoostRegressor())], XGBRegressor()

In the study, the hyperparameter settings of Z-score, IQR and MD methods, which are used to detect outliers and solve this problem, were determined, and analyzed. For the Z-score method, the "threshold" hyperparameter, which determines how many standard deviations away a data is from the standard deviation, is set to 3.0. This setting means that data that are more than 3 standard deviations away will be considered abnormal. For the IQR method, the "k" hyperparameter, which determines the distance between the upper and lower quartiles, is set to 1.5. For the MD method, the "threshold" hyperparameter, which determines whether the data are anomalous according to the MD distribution, is set to 95.0 percentile. This setting means that data with MD greater than a certain percentile will be considered outliers. These hyperparameter settings had a significant impact on model performance in anomaly detection and data preprocessing. The hyperparameter settings are presented in Table 3.

Table 3. Hyperparameter settings for outliers' treatment methods

Method	Hyperparameter	Settings
Z-score	threshold	3.0
IQR	k	1.5
MD	threshold	95.0 percentile

Python programming language was used for data analysis and model testing. In the data analysis process, pandas and NumPy libraries were used for data processing and manipulation. For model building and testing, the scikit-learn library was preferred. This library provides various ML algorithms and model evaluation tools. All processes were carried out in the Jupyter Notebook development environment, where code, text and visuals are presented together. A PC running on a Ryzen 7800x3D processor with a processor speed of 4.2 GHz was used for training the models. In addition, the PC has NVIDIA 4070 Ti GPU and 32 gigabyte 6000 MHz DDR5 RAM. Windows 11 was used as the operating system. The overview of the CO and NOx emission prediction system realized in the study is given in Figure 6.

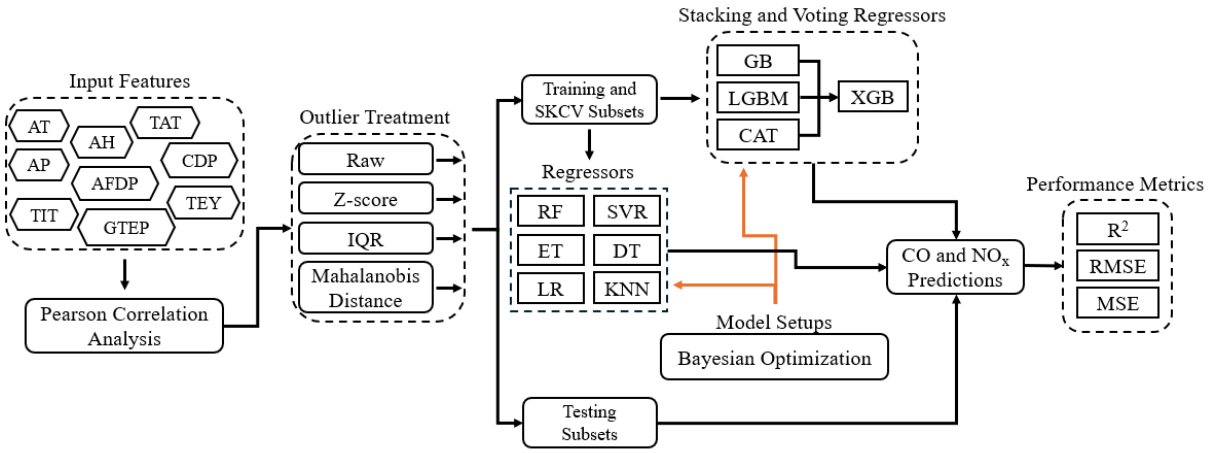

Figure 6. Overview of the CO and NOx emissions prediction system

3. Experimental Study and Results

In this section, the performance results of the regression models are presented and discussed. Table 4 shows the NOx emission prediction performance of the regression models and Table 5 shows the CO emission prediction performance of the models. The results are presented using three different metrics (R^2, RMSE, MSE) and four different outliers treatment methods (Raw, Z-Score, IQR, MD). R^2 indicates the explanatory power of the model, while RMSE and MSE indicate the error rates. The values of the best performing models for each outlier's treatment method are expressed in bold font. Table 4 shows that when the outliers in the dataset are treated with the MD method, ML models and ensemble methods reach the highest performance values in NOx emission prediction. In fact, the Stacking Regressor (SR) method reached the highest determination coefficient with $R^2 = 0.9974$ (RMSE = 0.5906, MSE = 0.3488). The predictions made by the SR model for NOx emissions closely match the observed (actual) NOX values, suggesting a strong agreement between the model's predictions and the real-world data. Figure 7 presents the R^2 scores of regression models for predicting NOx emissions, providing a clear comparison of their performance across different outlier treatment methods.

Upon reviewing the Table 4, it appears that the ET and DT models achieved an $R^2 = 1.0$, suggesting perfect performance compared to the SR. However, the respective RMSE values of 9.4325 and 1.5509 indicate that these models exhibit significantly larger prediction errors than expected. This discrepancy suggests that the models may have overfitted the training data, demonstrating excellent fit to the training set while lacking the ability to generalize to new data. Additionally, the MD method's sensitivity to data distribution means that outlier values in the dataset could negatively impact model performance, contributing to these observed errors. There is another very important point to be considered here. Figure 9 displays the comparison between predicted and actual NOx and CO emissions using the MD method and the SR model. The left panel plots the predicted NOx values against the actual NOx values. When the NOx predictions in Figure 9 are examined, it is seen that the points are ideally concentrated on the y = x line. This shows that the model predicts NOx values almost perfectly. These near-perfect predictions for NOx suggest that the model might be overfitting. A model that fits the training data exceptionally well may not maintain this performance when faced with new or noisier data. Analyzing the graphs and performance indicators reveals that the models trained on the dataset created using the MD method exhibit overfitting, indicating that the MD method does not yield accurate results for NOx prediction in the context of this study.

When Table 4 is further analyzed, the model created with SR in the dataset processed for outliers using the IQR method achieved the highest coefficient of determination ($R^2 = 0.9194$) and the lowest error values (RMSE = 2.7669, MSE = 7.6562). When the scatter plot in Figure 10 is examined, it is understood that the performance values obtained by the SR method seem to be suitable for real world data and the possibility of overfitting the model is low. In Figure 11, Hydrographs are given for each model trained on the dataset created with the IQR method and the prediction performances of the models are revealed. The success of the IQR method and the model built with SR in predicting NOx emissions at GT compared to other methods and models is clearly seen in the graphs. In addition to all this, it is found that the models trained on the dataset where outliers are processed with the Z-score method perform lower than the models trained on raw data.

Table 4. NOx emission prediction performance of regression models with outliers' treatment methods

	Raw			Z-Score		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8765	4.0443	16.3566	0.8716	4.1235	17.0035
ET	0.8908	3.8037	14.4683	0.8875	3.8607	14.9055
LR	0.4946	8.1832	66.9648	0.4911	8.2112	67.4251
SVR	0.7574	5.6698	32.1475	0.0739	11.0776	122.714
DT	0.7367	5.9059	34.8805	0.7383	5.8882	34.6718
KNN	0.8549	4.3846	19.2255	0.7893	5.2839	27.9197
VR	0.8575	4.3447	18.8766	0.8159	4.9384	24.3883
SR	**0.8942**	**3.7430**	**14.0101**	**0.8889**	**3.8353**	**14.7100**
	IQR			MD		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8992	3.0947	9.5776	0.9822	1.5602	2.4343
ET	0.9162	2.8218	7.9629	1.0	9.4325	8.8972
LR	0.7259	5.1046	26.0579	0.5233	8.0913	65.4694
SVR	0.0748	9.3789	87.9644	0.0788	11.2477	126.5129
DT	0.7672	4.7047	22.1351	1.0	1.5509	2.4052
KNN	0.8191	4.1473	17.2002	0.8723	4.1864	17.5262
VR	0.8511	3.7624	14.1559	0.8836	3.9972	15.9782
SR	**0.9194**	**2.7669**	**7.6562**	**0.9974**	**0.5906**	**0.3488**

Table 5 shows that ML models and ensemble methods achieve the highest performance values in CO emission prediction when outliers in the dataset are processed using the MD method. Similarly, examining the CO predictions in Figure 9, it is observed that the points are concentrated around the y = x line, but there is a more pronounced scatter and deviations. Although the CO predictions are more modest, there is a high probability that the model may have overfitted. Therefore, when Table 5 is further analyzed using other methods, it is seen that the model created with SR reaches the highest coefficient of determination (R^2 = 0.8556) and the lowest error values (RMSE = 0.4619, MSE = 0.2133) in the dataset processed for outliers using the IQR method. The scatter plot in Figure 10 also supports these results obtained by SR with the IQR method. In addition, the Hydrographs given for each model in Figure 12 reveal the high success of the IQR method and the model built with SR in predicting CO emissions in GT compared to other methods and models. Also, similar to the results for NOx emission prediction, the CO prediction performance of the models trained on raw data is lower than the models trained on the dataset generated by the IQR method, but higher than the scenario using the Z-score method. Figure 8 displays the R^2 scores of regression models for CO emission prediction, offering a concise comparison of their performance across various outlier treatment methods.

Table 5. CO emission prediction performance of regression models with outliers' treatment methods

	Raw			Z-Score		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.7649	1.1306	1.2783	0.7509	1.1637	1.3543
ET	0.7990	1.0453	1.0926	**0.7682**	**1.1225**	**1.2602**
LR	0.5566	1.5526	2.4107	0.5511	1.5621	2.4404
SVR	0.6755	1.3282	1.7641	0.4405	1.7441	3.0419
DT	0.5081	1.6354	2.6747	0.4055	1.7978	3.2321
KNN	0.7630	1.1350	1.2883	0.6626	1.3543	1.8343
VR	0.7631	1.1349	1.2880	0.6220	1.4335	2.0550
SR	**0.8026**	**1.0359**	**1.0731**	0.6705	1.3384	1.7914
	IQR			MD		
	R^2	RMSE	MSE	R^2	RMSE	MSE
RF	0.8426	0.4822	0.2325	0.9682	0.4002	0.1602
ET	0.8542	0.4640	0.2153	1.0	4.4572	1.9867
LR	0.6150	0.7542	0.5689	0.5651	1.4803	2.1915
SVR	0.5491	0.8163	0.6664	0.4658	1.6408	2.6922
DT	0.6658	0.7027	0.4939	1.0	3.1728	1.0066
KNN	0.7752	0.5764	0.3322	0.7983	1.0080	1.0162
VR	0.8060	0.5353	0.2866	0.9286	0.5998	0.3597
SR	**0.8556**	**0.4619**	**0.2133**	**0.9974**	**0.1122**	**0.0126**

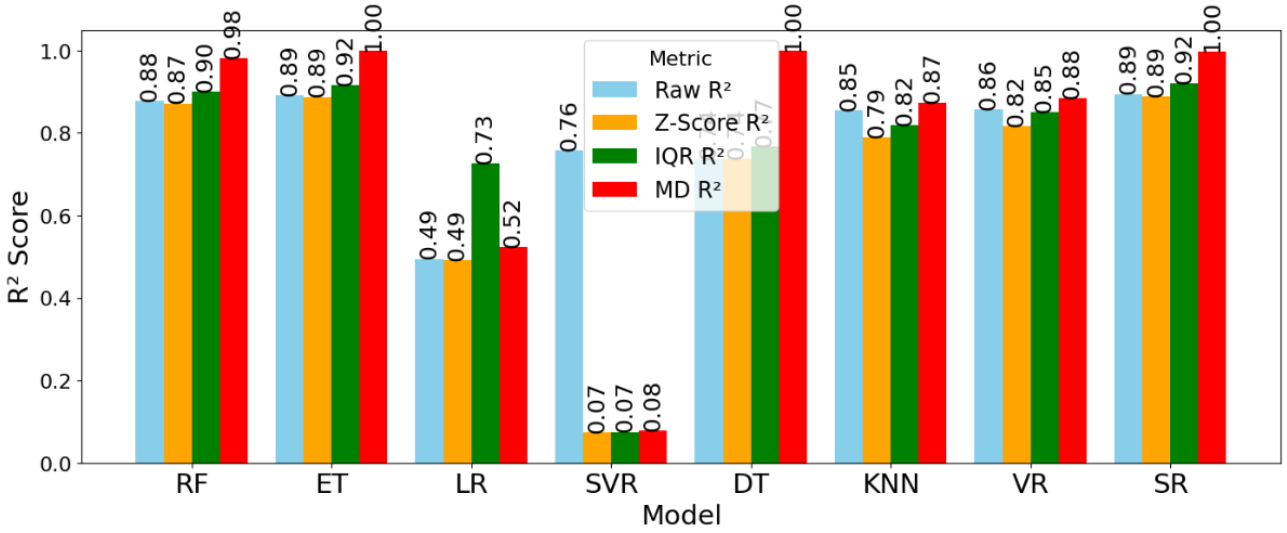

Figure 7. Comparison of R^2 scores for NOx emission prediction models with different outlier treatment methods

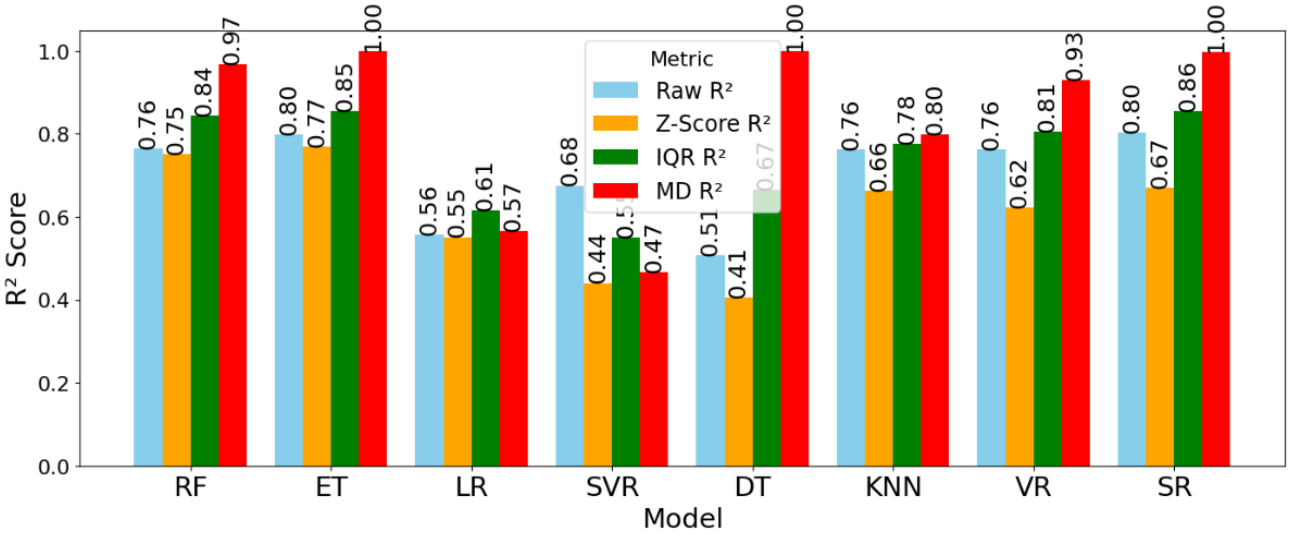

Figure 8. Comparison of R^2 scores for CO emission prediction models with different outlier treatment methods

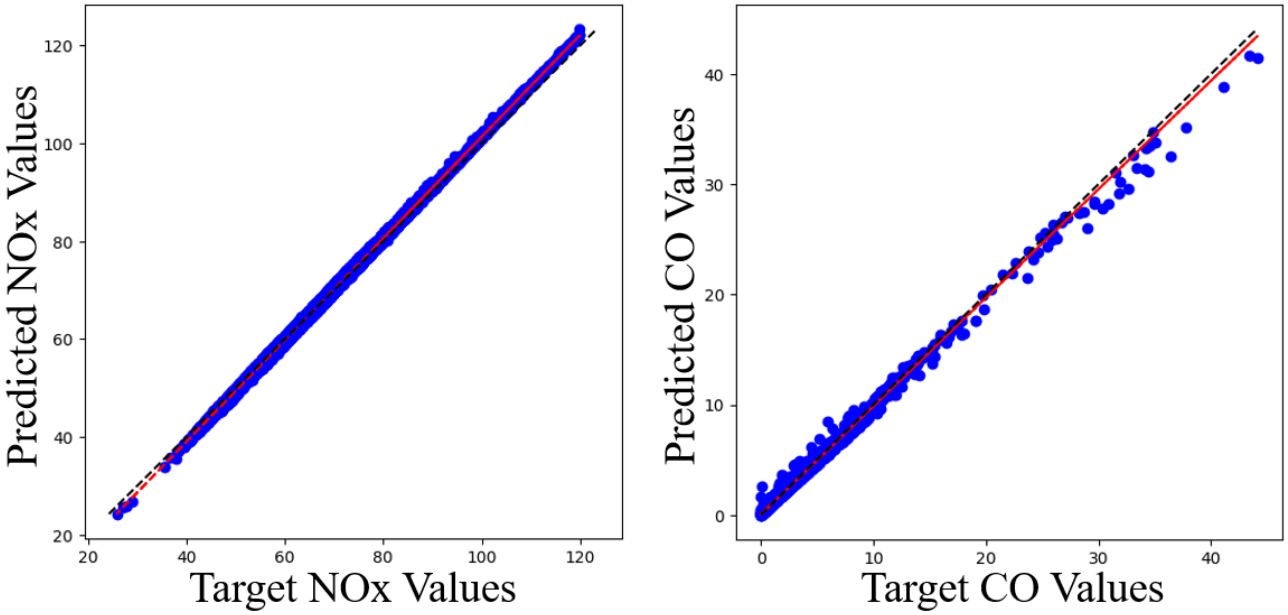

Figure 9. Scatter plots generated on the MD method for the SR prediction model

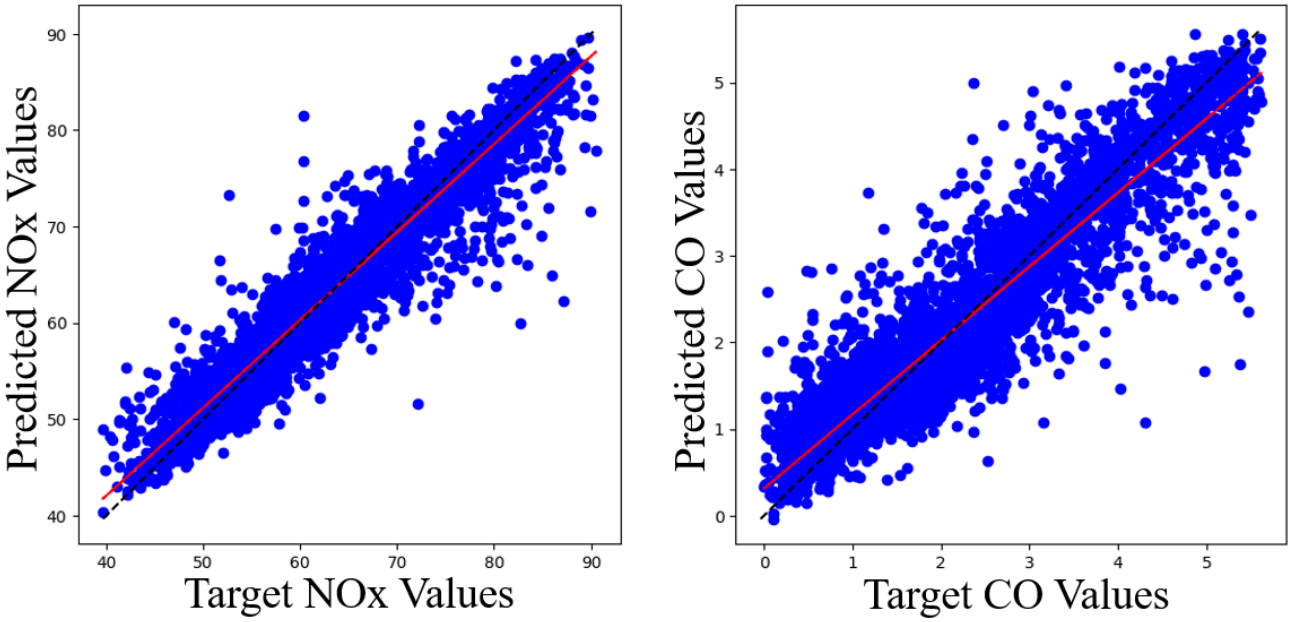

Figure 10. Scatter plots generated on the IQR method for the SR prediction model

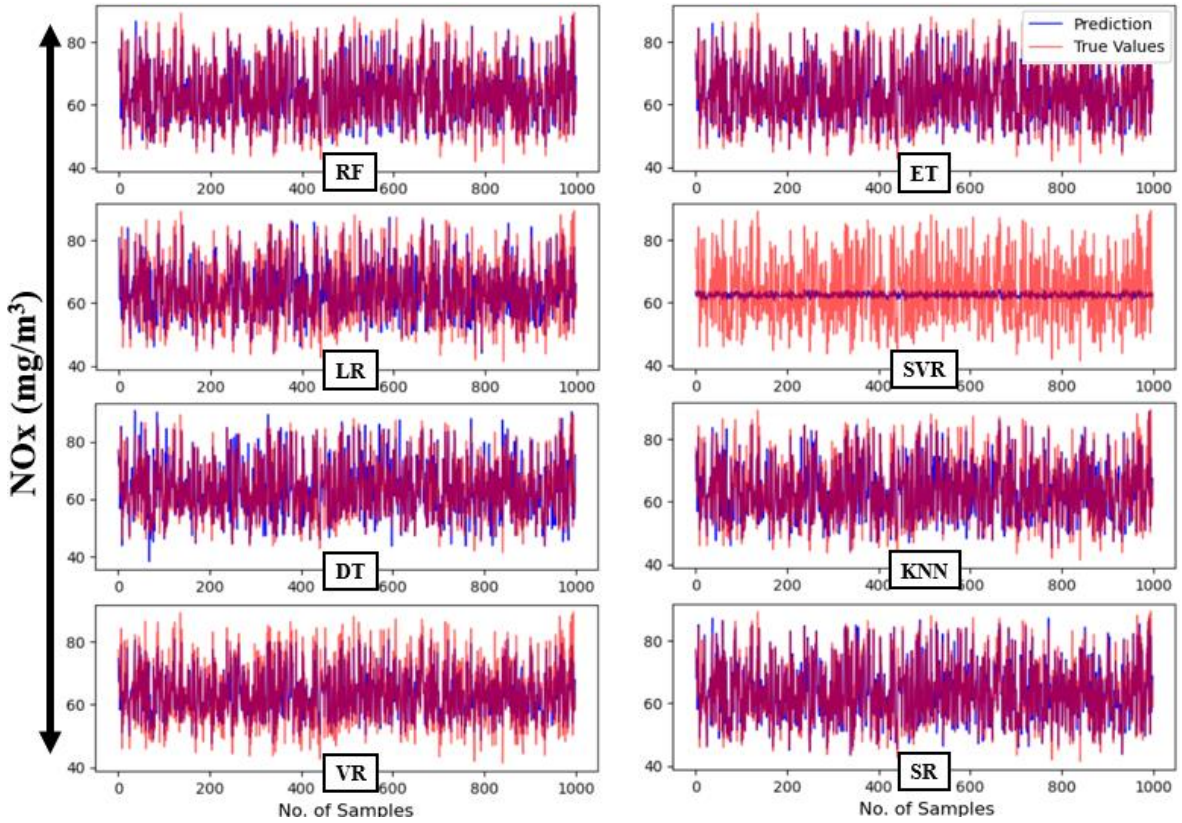

Figure 11. Hydrographs for all designed ML models for NOx in the IQR-processed dataset

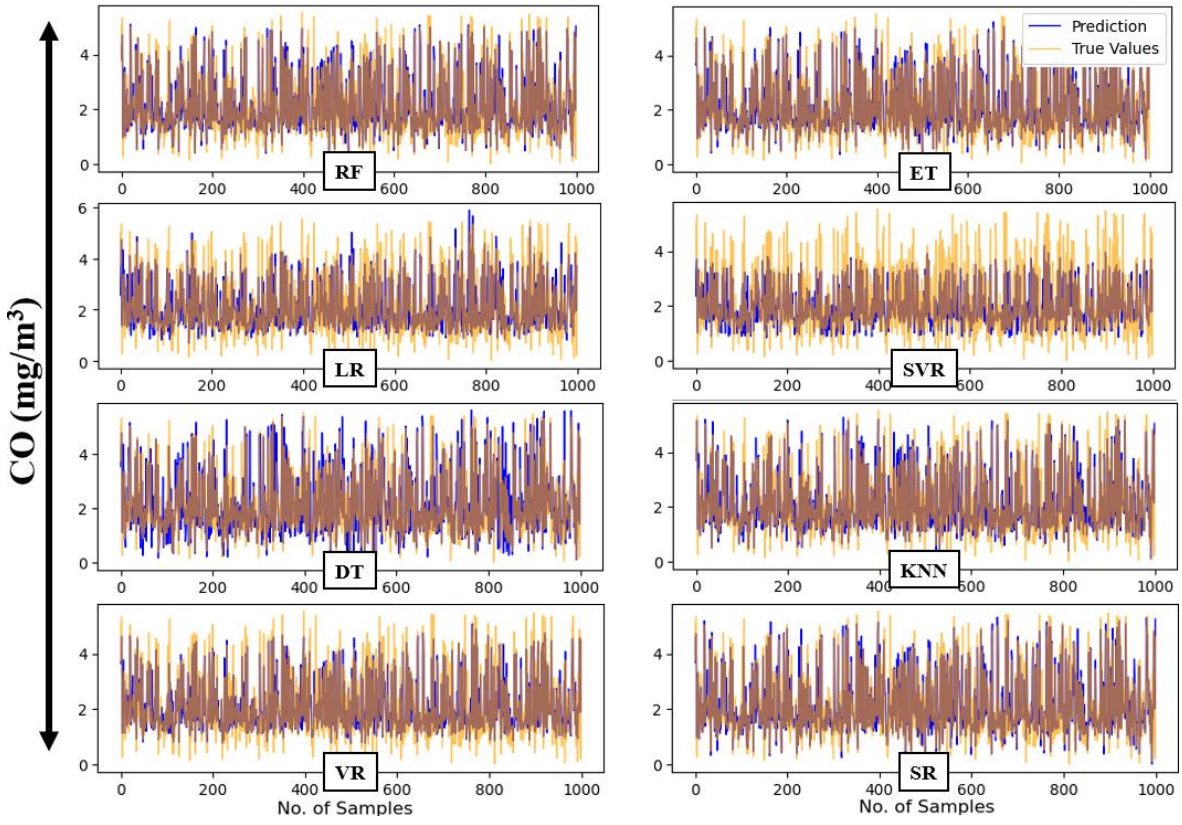

Figure 12. Hydrographs for all designed ML models for CO in the IQR-processed dataset

Figure 13 shows the results of the sensitivity analysis for CO and NOx emissions. The graph reflects the impact of each feature on the predictive performance of the model. Blue bars represent CO sensitivity and red bars represent NOx sensitivity. The sensitivity analysis for each feature is performed as follows:

1. Baseline score: First, the error of the model's predictions (MSE) with the available features is calculated (Equation 16).
2. Perturbed scores: For each feature, the error of the model's predictions is recalculated by randomly permuting the values of this feature. This process is repeated 10 times and the MSE values obtained each time are recorded (Equation 17). Here $\hat{y}_j^{(m)}$, denotes the values predicted by the model after the m-th permutation and m represents the number of permutations (10).

$$\text{Perturbed\_Score}^{(m)} = \frac{1}{n}\sum_{j=1}^{n}\left(y_i - \hat{y}_j^{(m)}\right)^2 \quad (17)$$

3. Sensitivity value: The sensitivity value for each feature is calculated by subtracting the average error from the permuted scores from the baseline score (Equation 18).

$$\text{Sensitivity}(i) = \left(\frac{1}{10}\sum_{m=1}^{10}\text{Perturbed\_Score}^{(m)}\right) - \text{Baseline\_Score} \quad (18)$$

When examining Figure 13, it is found that AT has a significant impact on CO emissions, with a sensitivity value of 0.6234. It also shows a notable effect on NOx emissions (2.5019). Higher temperatures generally enhance combustion, promoting NOx formation while potentially reducing CO emissions. AP has a low-level effect on CO and a moderate impact on NOx emissions. The impact of AP on CO is measured at 0.0587, while its impact on NOx is 0.7457. TIT has the highest impact on CO emissions, with a sensitivity value of 10.0492. Its effect on NOx is negative, with a sensitivity value of -0.6029. The high sensitivity of TIT on CO emissions indicates that TIT plays a critical role in combustion efficiency and, consequently, CO formation. The negative effect on NOx emissions may indicate that high inlet temperatures can reduce NOx formation. TAT has a significant impact on CO emissions, with a sensitivity value of 6.819. Its effect on NOx is quite low and positive, measured at 0.0694. The high impact of TAT on CO emissions suggests that the turbine outlet temperature can affect the composition of post-combustion gases. Overall, it is understood that TIT and TAT parameters are particularly critical in predicting CO emissions, while AT has significant effects on NOx emissions.

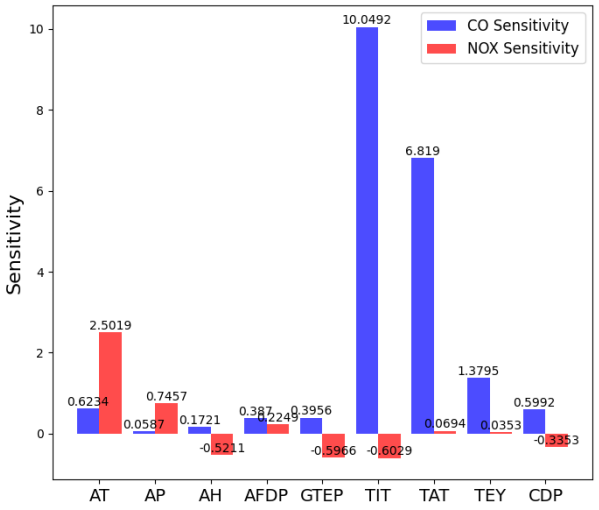

Figure 13. Parameter sensitivity analysis results for CO and NOx emissions

The error deviation graphs presented in Figure 14 show the differences between the predicted and actual CO and NOx emission rates. Table 6 provides the minimum and maximum error values obtained for each predictive ML model created. For the SR model trained on the dataset processed with the IQR method for outlier treatment, the minimum and maximum error values for NOx are -21.5874 and 24.5953, respectively, while for CO they are -2.6273 and 3.5575, respectively. Compared to other models, it can be concluded that the SR model has the lowest error deviation. This indicates that the SR model is more successful in predicting CO and NOx emissions and that these predictions are closer to the actual values.

Table 6. Error deviation for all designed predictive models in the IQR-processed dataset

Error deviation	RF	ET	LR	SVR	DT	KNN	VR	SR
NOx								
Min deviation	-27.1536	-21.9554	-27.3576	-21.9011	-34.0320	-25.9326	-18.3893	-21.5874
Max deviation	30.1443	24.4523	33.1878	28.0568	37.6850	31.1432	26.4236	24.5953
CO								
Min deviation	-2.7295	-2.6710	-3.3075	-2.8513	-4.7361	-2.9963	-2.9103	-2.6273
Max deviation	3.7549	3.6115	4.0626	3.8812	4.6908	3.5781	3.4712	3.5575

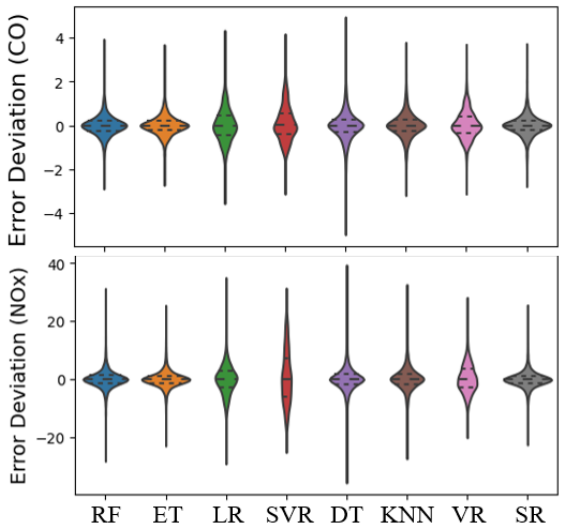

Figure 14. Error deviation comparison of all predictive models for NOx and CO in the IQR-processed dataset

4. Discussion

This study aims to develop predictive models for predicting CO and NOx emissions from GT using various ML algorithms and compare their performance with VR and SR methods. A comprehensive evaluation of these models is conducted to provide insights into their practical applicability in industrial settings. The findings show the importance of hyperparameter tuning and outlier processing in improving the prediction accuracy of these models.

Tree-based methods using a combination of models provided better performance compared to other regression models for both CO and NOx predictions. SR achieved the highest determination coefficient values when outliers in the dataset were treated using the MD method. However, despite the high R^2 values, high RMSE values indicate the possibility of overfitting, mainly in NOx predictions. While the MD method effectively detects outliers, it can lead to overfitting if the dataset does not follow a normal distribution. Overfitting occurs when the model performs well on training data but poorly on new data (Karthikeyan et al. 2023). Furthermore, the risk of overfitting associated with the MD method, as observed in this study, has also been noted in prior research. For instance, Ghorbani (2019) highlighted that the MD method, while effective for multivariate outlier detection, can be overly sensitive to data distribution and may lead to overfitting in small or homogeneous datasets. This corroborates our findings, where the MD method achieved high R^2 values but showed poor generalization in real-world scenarios.

In contrast, the IQR method's robustness and consistency make it a more suitable choice for practical applications in emission monitoring and control. In addition, the MD method considers the distance of each data point in the data set to other data points in multiple dimensions (Leys et al. 2018). This method can identify extreme data points as outliers and remove them from the model or cause the model to overfit these points. Another disadvantage of the MD method is that it is very sensitive to the distribution of the dataset (Todeschini et al. 2013). If there is heterogeneity or too much noise in the dataset, this method may detect false positive outliers. In this case, the model may focus on false positive outliers instead of true outliers, which weakens the generalization ability of the model. Furthermore, if there is too much variance in the dataset, the MD method may misinterpret this variance and flag some data points as outliers (Wu et al. 1997). This can cause the model to be trained on an incorrect subset of data and lead to overfitting.

On the other hand, the IQR method for outlier treatment produced more robust models with lower error rates. The SR model trained on the IQR-processed dataset achieved the highest determination coefficient ($R^2 = 0.9194$) and the lowest RMSE and MSE values for NOx predictions. The scatter plots in Figure 10 confirm the model's performance close to the true values, indicating a low risk of overfitting. This aligns with previous research by Yaro et al. (2023), who demonstrated that the IQR method is highly effective in reducing the impact of outliers in datasets with non-normal distributions, leading to more robust and generalizable models. Similarly, Mishra et al. (2019) emphasized that the IQR method outperforms Z-Score and MD in scenarios where data variability is high, as it is less sensitive to outlier values and provides a balanced approach to outlier detection. These studies collectively support the conclusion that the IQR method is a reliable choice for emission prediction tasks, especially in datasets with complex operational variability, such as those from gas turbines.

Similar patterns were observed for CO predictions. Although the MD method initially seemed to give the best results, further analysis revealed significant spread and biases, indicating the possibility of overfitting. The SR model trained on the IQR processed dataset outperformed the other models, again achieving the highest R^2 (0.8556) and the lowest RMSE and MSE values. This observation supports that the IQR method offers a balanced approach to dealing with outliers and ensures the robustness and accuracy of the model.

The worse performance of models trained on raw data compared to models trained on the processed dataset emphasizes the necessity of outlier handling methods in improving the reliability of emission prediction models. This finding is in line with previous research (Osborne and Overbay, 2019), which indicates that untreated outliers can significantly skew model training and prediction results. However, in the dataset where outliers were treated with the Z-score method, the performance of the models decreased compared to the raw dataset. The Z-score method is a common method for detecting and treating outliers, but it is based on the assumption of a normal distribution. If the dataset is not normally distributed or is multidimensional, this method may detect false positive outliers and over-smooth the

variance in the dataset (Mare et al. 2017). This reduces the model's ability to capture true patterns and leads to performance degradation. The degradation of the performance of the models in the dataset processed with the Z-score method compared to the raw dataset indicates that this method is not always appropriate and should be applied carefully according to the characteristics of the dataset.

In the study, sensitivity analysis was performed to improve the performance of emission prediction models and to identify the parameters that have the most impact on emissions. The findings show that turbine inlet and outlet temperatures (TIT and TAT) have a significant effect on CO emissions. TIT has the highest impact on CO emissions, indicating that TIT plays a critical role on combustion efficiency and hence CO formation. The negative effect on NOx emissions indicates that higher inlet temperatures can reduce NOx formation. Similarly, AT has a significant effect on NOx emissions, but less on CO emissions. This finding suggests that higher temperatures can reduce CO emissions while promoting NOx formation.

In Table 7, the results obtained by using the SR model proposed in this study and the dataset processed with the IQR method are compared with other studies in literature. The table shows that the model proposed in this study outperforms other studies in the literature in the prediction of NOx and CO emissions. The RMSE value of the proposed model in NOx prediction is 2.76 and the RMSE value in CO prediction is 0.46. These error values are significantly lower than the results of other models in the literature. In addition, higher R^2 values were obtained in both NOx and CO emissions prediction compared to other models in the literature. Possible reasons for this high performance include the effective treatment of outliers with the IQR method, which reduces the noise in the dataset. The IQR method provides a robust approach to handling non-normally distributed data by eliminating extreme values without over-penalizing potential influential data points. Moreover, the structure of the SR model and hyperparameter optimization are other factors that positively affect the prediction performance. Although the MD method initially appeared to provide competitive results in terms of R^2 values, its over-performance can be attributed to its sensitivity to multivariate relationships and its assumption of normal data distribution. However, this sensitivity may lead to overfitting, as the MD method can classify influential but valid data points as outliers, thereby reducing model generalization. This explains why, despite high R^2 values, the RMSE values remained relatively high, indicating possible model overfitting and reduced performance on new data.

Table 7. Comparison of NOx and CO emission prediction models from the literature with the proposed SR

Model and Reference	NOx			CO		
	RMSE	MSE	R^2	RMSE	MSE	R^2
ANN with feature normalization (Nino-Adan et al. 2021)	7.06	-	0.57	-	-	0.43
Symbolic regression (Kochueva and Nikolskii, 2021)	-	-	0.83	-	-	0.89
KNN (Rezazadeh, 2021)	-	-	0.89	-	-	-
DFR (Coelho et al. 2024)	5.54	30.69	-	1.35	1.84	-
KNN (Wood, 2023)	5.12	-	-	-	-	-
ANFIS (Dirik, 2022)	4.98	24.8	-	-	-	-
SR (Pachauri, 2024)	3.83	14.70	0.87	0.61	0.37	0.77
This study (SR with IQR Treatment)	2.76	7.65	0.92	0.46	0.21	0.85

5. Conclusion

This study aims to evaluate the performance of various ML algorithms for predicting CO and NOx emissions from GT. The algorithms used in the study include RF, ET, LR, SVR, DT, KNN, as well as VR and SR methods. In the ensemble methods, GB, LightGBM and CatBoost algorithms were used as base learners and XGBoost was determined as the meta-learner.

The study examined the effects of processing the outliers in the dataset with various methods (Z-Score, IQR, MD) on model performance. The findings show that the MD method provides high performance of the models, especially in NOx emission prediction, but it also brings the risk of overfitting. Although the SR model provides high R^2 and low error values in NOx predictions, scatter plots and hydrographs reveal that the data processed with the MD method do not show a distribution suitable for real world data. Due to the sensitivity of the MD method to outlier values, some models tend to overfit, indicating that caution should be exercised in the use of this method.

The treatment of outliers with the IQR method provided more balanced and generalizable results in CO and NOx emission predictions. In the dataset processed with the IQR method, the SR model achieved R^2 values of 0.9194 in NOx emission predictions and 0.8556 in CO emission predictions and showed low error rates in other metrics. Scatter plots and hydrographs also showed a consistent distribution. These results show that the IQR method is an effective approach to deal with outliers and has an impact on improving the performance of ML models. In the dataset processed with the Z-Score method, the performance of the models was lower compared to the models trained with raw data. This finding suggests that the Z-Score method may be less effective in identifying and processing outliers,

especially in this dataset, and that each outlier processing method may yield different results depending on the characteristics of the dataset and the problem definition.

The results show that the methods used to handle outliers have a significant impact on the performance of ML models and that the right choice of method can improve model accuracy and generalizability. While the IQR method gives balanced results in predicting CO and NOx emissions, the MD method can provide high performance in some cases, although it increases the risk of overfitting. ML algorithms and outlier treatment methods need to be carefully selected for gas turbine emission prediction.

The sensitivity analysis highlighted the significant impact of TIT, TAT, and AT on CO and NOx emissions in the prediction models. In particular, the high impact of TIT on CO emissions shows how important a role TIT plays on combustion efficiency and CO formation. This highlights the need for careful control of TIT to optimize combustion processes and reduce CO emissions. Furthermore, the negative effect of TIT on NOx emissions indicates that high inlet temperatures can reduce NOx formation. This finding suggests that high temperature processes in power generation have the potential to control NOx emissions. The significant effect of AT on NOx emissions suggests that environmental temperature conditions should also be considered in emission control strategies.

Based on the results, some recommendations can be made to improve gas turbine emission predictions and develop environmental management strategies in the energy sector. First of all, the quality of sensors and data collection systems in power plants should be improved and data accuracy should be ensured through regular calibrations, as accurate detection and processing of outliers directly affects model performance. The use of integrated model approaches should be encouraged as combinations of different ML algorithms can provide more balanced and accurate results in emission predictions. In addition, it is understood that temperature parameters are critical for improving the accuracy of emission prediction models and obtaining reliable predictions. To improve emissions management in the power sector, turbine and environmental temperatures need to be optimized. These approaches will enable more effective implementation of emission reduction strategies and contribute to reducing environmental impacts. To increase the generalizability of the results, future studies should test these methods on different datasets or systems, ensuring their applicability across various operational contexts and conditions.

The findings of this study hold significant potential to support sustainable energy policies and environmental management practices. For instance, the high accuracy of the SR model can help Turkish industries comply with stringent emission regulations, such as the EU Industrial Emissions Directive, while aligning with global climate goals like the Paris Agreement. By enabling precise real-time emission monitoring, these models can empower policymakers to design data-driven regulations and incentivize the adoption of low-emission technologies. Furthermore, optimizing TIT based on sensitivity analysis insights could reduce CO emissions by up to 10–15% in practical scenarios, contributing to cleaner air and improved public health in countries. Integrating ML-driven outlier detection and ensemble methods into existing emission control systems can also reduce operational costs by minimizing fuel waste and avoiding non-compliance penalties. These advancements not only enhance the sustainability of gas turbine operations but also position the countries' energy sector to meet evolving environmental standards while maintaining energy security and supporting its transition to a greener economy.

References

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production, 203*, 810-821. https://doi.org/10.1016/j.jclepro.2018.08.207

Aslan, E. (2024). Prediction and Comparative Analysis of Emissions from Gas Turbines Using Random Search Optimization and Different MachineL earning Based Algorithms. *Bulletin of the Polish Academy of Sciences Technical Sciences*, e151956-e151956. https://doi.org/10.24425/bpasts.2024.151956

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., ... & Carpenter, A. E. (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods*, *14*(9), 849-863.

Coelho, D. S. L., Ayala, H. V. H., & Mariani, V. C. (2024). CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering. *Fuel*, *355*, 129366. https://doi.org/10.1016/j.fuel.2023.129366

Dalal, A. S., Sultanova, N., Jayabalan, M., & Mustafina, J. (2023, December). Gas turbine–CO & NOx emission data analysis with predictive modelling using ML/AI approaches. In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 100-104). IEEE. https://doi.org/10.1109/DeSE60595.2023.10469322

Dirik, M. (2022). Prediction of NOx emissions from gas turbines of a combined cycle power plant using an ANFIS model optimized by GA. *Fuel*, *321*, 124037. https://doi.org/10.1016/j.fuel.2022.124037

Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., & Torres, J. F. (2018). Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, *11*(4), 949. https://doi.org/10.3390/en11040949

Farzaneh-Gord, M., & Deymi-Dashtebayaz, M. (2011). Effect of various inlet air cooling methods on gas turbine performance. *Energy*, *36*(2), 1196-1205. https://doi.org/10.1016/j.energy.2010.11.027

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583-595. https://doi.org/10.22190/FUMI1903583G

Karthikeyan, S., Kathirvalavakumar, T., & Prasath, R. (2023, June). Classification of the Class Imbalanced Data Using Mahalanobis Distance with Feature Filtering. In *International Conference on Mining Intelligence and Knowledge Exploration* (pp. 45-53). Cham: Springer Nature Switzerland.

Kaya, H., Tüfekci, P., & Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences*, *27*(6), 4783-4796. https://doi.org/10.3906/elk-1807-87

Kochueva, O., & Nikolskii, K. (2021). Data analysis and symbolic regression models for predicting CO and NOx emissions from gas turbines. *Computation*, *9*(12), 139. https://doi.org/10.3390/computation9120139

Kumar, M. V., Babu, A. V., Reddy, C. R., Pandian, A., Bajaj, M., Zawbaa, H. M., & Kamel, S. (2022). Investigation of the combustion of exhaust gas recirculation in diesel engines with a particulate filter and selective catalytic reactor technologies for environmental gas reduction. *Case Studies in Thermal Engineering*, *40*, 102557.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150-156. https://doi.org/10.1016/j.jesp.2017.09.011

Liu, Z., Meng, H., Huang, J., Kwangwari, P., Ma, K., Xiao, B., & Li, L. (2021). Acute carbon monoxide poisoning with low saturation of carboxyhaemoglobin: a forensic retrospective study in Shanghai, China. *Scientific Reports*, *11*(1), 18554.

Lopes, C., Antelo, L. T., Franco-Uría, A., Alonso, A. A., & Pérez-Martín, R. (2015). Valorisation of fish by-products against waste management treatments–Comparison of environmental impacts. *Waste Management*, *46*, 103-112. https://doi.org/10.1016/j.wasman.2015.08.017

Lott, P., Casapu, M., Grunwaldt, J. D., & Deutschmann, O. (2024). A review on exhaust gas after-treatment of lean-burn natural gas engines–From fundamentals to application. *Applied Catalysis B: Environmental*, *340*, 123241.

Mahmoudi, S., Baeyens, J., & Seville, J. P. (2010). NOx formation and selective non-catalytic reduction (SNCR) in a fluidized bed combustor of biomass. *Biomass and bioenergy*, *34*(9), 1393-1409. https://doi.org/10.1016/j.biombioe.2010.04.013

Mare, D. S., Moreira, F., & Rossi, R. (2017). Nonstationary Z-score measures. *European Journal of Operational Research*, *260*(1), 348-358. https://doi.org/10.1016/j.ejor.2016.12.001

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147. https://doi.org/10.38094/jastt1457

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67-72. https://doi.org/10.4103/aca.ACA_157_18

Naghibi, A. (2024). Exploring explainable ensemble machine learning methods for long-term performance prediction of industrial gas turbines: A comparative analysis. *Engineering Applications of Artificial Intelligence*, *138*, 109318. https://doi.org/10.1016/j.engappai.2024.109318

Nino-Adan, I., Portillo, E., Landa-Torres, I., & Manjarres, D. (2021). Normalization influence on ANN-based models performance: A new proposal for Features' contribution analysis. *IEEE Access*, *9*, 125462-125477. https://doi.org/10.1109/ACCESS.2021.3110647

Osborne, J. W., & Overbay, A. (2019). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6. https://doi.org/10.7275/qf69-7k43

Pachauri, N. (2024). An emission predictive system for CO and NOx from gas turbine based on ensemble machine learning approach. *Fuel*, *366*, 131421. https://doi.org/10.1016/j.fuel.2024.131421

Pandey, R. A., & Chandrashekhar, B. (2014). Physicochemical and biochemical approaches for treatment of gaseous emissions containing NOx. *Critical Reviews in Environmental Science and Technology*, *44*(1), 34-96. https://doi.org/10.1080/10643389.2012.710430

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, 972421. https://doi.org/10.3389/fnano.2022.972421

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

Rezazadeh, A. (2021). Environmental pollution prediction of NOx by predictive modelling and process analysis in natural gas turbine power plants. *Pollution*, *7*(2), 481-494. https://doi.org/10.22059/poll.2021.316327.977

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, *251*, 26-34. https://doi.org/10.1016/j.neucom.2017.04.018

Tian, J., Wang, L., Xiong, Y., Wang, Y., Yin, W., Tian, G., ... & Ji, S. (2024). Enhancing combustion efficiency and reducing nitrogen oxide emissions from ammonia combustion: A comprehensive review. *Process Safety and Environmental Protection*, 183, 514-543. https://doi.org/10.1016/j.psep.2024.01.020

Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., & Filzmoser, P. (2013). Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection. *Analytica Chimica Acta*, *787*, 1-9. https://doi.org/10.1016/j.aca.2013.04.034

Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754-757.

Wardana, M. K. A., & Lim, O. (2022). Review of improving the NOx conversion efficiency in various diesel engines fitted with SCR system technology. *Catalysts*, *13*(1), 67.

Wood, D. A. (2023). Long-term atmospheric pollutant emissions from a combined cycle gas turbine: Trend monitoring and prediction applying machine

learning. *Fuel*, *343*, 127722. https://doi.org/10.1016/j.fuel.2023.127722

Wu, T. J., Burke, J. P., & Davison, D. B. (1997). A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, 1431-1439. https://doi.org/10.2307/2533509

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Yaro, A. S., Maly, F., & Prazak, P. (2023). Outlier detection in time-series receive signal strength observation using z-score method with s n scale estimator for indoor localization. *Applied Sciences*, *13*(6), 3900. https://doi.org/10.3390/app13063900

Yousif, S. T., Ismail, F. B., & Al-Bazi, A. (2024). A hybrid neural network-based improved PSO algorithm for gas turbine emissions prediction. *Advanced Theory and Simulations*, 2301222. https://doi.org/10.1002/adts.202301222

Yu, H., & Kim, S. (2012). SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural Computing*, *1*, 479-506.