# การพัฒนาแบบจำลองเพื่อพยากรณ์การปล่อยไนโตรเจนออกไซด์ (NOx) ของโรงไฟฟ้าน้ำพองด้วย Machine Learning

# Development of Nitrogen Oxides (NOx) Emission Prediction Model of Nam Phong Power Plant with Machine Learning

วิสิษฐ์ ธีระวงศ์[1] นที พนากานต์[2*] สุจินต์ บุรีรัตน์[2] ณัญธิวัฒน์ พลดี[2] และ ชนกนันท์ สุขกำเนิด[3]
Wisit Teerawong[1], Natee Panagant[2*], Sujin Bureerat[2], Nantiwat Pholdee[2]
and Chanoknun Sookkumnerd[3]
Received: 1 November 2023, Revised: 27 December 2023, Accepted: 2 January 2024

## บทคัดย่อ

ก๊าซไนโตรเจนออกไซด์ (Nitrogen Oxides : NOx) เป็นกลุ่มก๊าซที่เป็นอันตรายต่อสุขภาพและสิ่งแวดล้อม โดยมีแหล่งกำเนิดหลักจากการเผาไหม้เชื้อเพลิงในเครื่องยนต์และกระบวนการอุตสาหกรรม ดังนั้นเพื่อปฏิบัติตามข้อกำหนด โรงไฟฟ้าน้ำพองจึงติดตั้งเครื่องตรวจวัดมลพิษทางอากาศจากปล่องแบบอัตโนมัติอย่างต่อเนื่อง (Continuous Emission Monitoring System : CEMS) และรายงานค่าไปยังหน่วยงานที่เกี่ยวข้อง อย่างไรก็ตามระบบ CEMS เป็นระบบที่มีความซับซ้อนมีค่าใช้จ่ายสูง ประกอบกับในปี พ.ศ. 2565 ประเทศไทยได้ประกาศใช้ข้อกำหนดใหม่ที่อนุญาตให้รายงานค่า NOx โดยวิธีคาดคะเนแทนระบบ CEMS ได้ จึงได้นำมาสู่การศึกษาและพัฒนาแบบจำลองเพื่อพยากรณ์การปล่อย NOx ของโรงไฟฟ้าน้ำพองด้วย Machine Learning โดยในงานวิจัยนี้ได้ศึกษาเปรียบเทียบประสิทธิภาพแบบจำลองที่แตกต่างกันจำนวน 6 Algorithms ได้แก่ Linear Regression, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors และ Backpropagation Multilayer Perceptron Neural Network ซึ่งพบว่าแบบจำลองของ Random Forest มีประสิทธิภาพการพยากรณ์ที่สูงกว่าแบบจำลองอื่น ๆ โดยมีค่า MAE และ MAPE ต่ำสุด และค่า $R^2$ สูงสุด อีกทั้งการศึกษานี้ยังพบว่าอุณหภูมิไอน้ำที่ฉีดเข้าห้องเผาไหม้ของเครื่องกังหันก๊าซเป็นพารามิเตอร์ที่สำคัญต่อความแม่นยำของแบบจำลองและมีผลต่อการเกิด NOx ซึ่งสามารถใช้เป็นแนวทางควบคุมหรือลดปริมาณการปล่อย NOx โดยไม่กระทบต่อประสิทธิภาพของโรงไฟฟ้า

**คำสำคัญ:** พยากรณ์การปล่อย NOx, Machine Learning, โรงไฟฟ้าน้ำพอง, CEMS, PEMS

---

[1] โรงไฟฟ้าน้ำพอง (กฟผ.) อำเภอน้ำพอง จังหวัดขอนแก่น 40310

[1] Nam Phong Power Plant (EGAT), Nam Phong, Khon Kaen 40310, Thailand.

[2] ศูนย์วิจัยและพัฒนาโครงสร้างมูลฐานอย่างยั่งยืน ภาควิชาวิศวกรรมเครื่องกล คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น อำเภอเมือง จังหวัดขอนแก่น 40002

[2] Sustainable Infrastructure Research and Development Center, Department of Mechanical Engineering, Faculty of Engineering, Khon Kaen University, Muang, Khon Kaen 40002, Thailand.

[3] ภาควิชาวิศวกรรมเครื่องกล คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น อำเภอเมือง จังหวัดขอนแก่น 40002

[3] Department of Mechanical Engineering, Faculty of Engineering, Khon Kaen University, Muang, Khon Kaen 40002, Thailand.

[*] Corresponding author, e-mail: natepa@kku.ac.th

# ABSTRACT

Nitrogen Oxides (NOx) are harmful gases to human health and the environment. These emissions primarily result from fuel combustion in engines and industrial processes. To meet regulatory requirements, the Nam Phong Power Plant in Thailand has implemented Continuous Emission Monitoring Systems (CEMS) to measure and report NOx emissions to regulatory authorities. However, considering the high costs associated with installing and maintaining CEMS, as well as recent changes in Thai legislation allowing for predictive NOx measurement methods, it is worth exploring the use of Machine Learning as a reliable method for estimating NOx emissions accurately. In this study, a comprehensive comparison was conducted on six Machine Learning algorithms: Linear Regression, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors, and Backpropagation Multilayer Perceptron Neural Network. Among these models, Random Forest emerged as the top performer, exhibiting superior performance metrics, including the lowest MAE, MAPE, and the highest $R^2$ scores. These results underscore the potential accuracy and reliability of Random Forest in predicting NOx emissions. Furthermore, research on feature importance has revealed the significant influence of certain parameters on model accuracy. These parameters include steam injection flow, steam injection temperature, and ambient conditions. The influence of controllable factors, such as the temperature of steam injection, on NOx emissions is noteworthy. These findings not only hold promise for enhancing the precision of predictive models but also present opportunities to decrease NOx emission levels while maintaining plant efficiency.


**Key words:** NOx prediction, machine learning, Nam Phong power plant, CEMS, PEMS

## INTRODUCTION

Nitrogen Oxides (NOx) are harmful gases that have been identified as major contributors to air pollution, with negative impacts on both the environment and human health (United States Environmental Protection Agency, 2023). The primary sources of NOx include fuel combustion in engines and manufacturing processes. To comply with regulations, power plants in Thailand, including the Nam Phong Power Plant, have installed Continuous Emission Monitoring Systems (CEMS) to measure and report their NOx emissions to regulatory authorities (Department of Industrial Works, 2007). However, CEMS is a complex and expensive system, which has led to a search for alternative methodologies. Additionally, the introduction of new environmental regulations in Thailand (Ministry of Industry, 2022) enables the utilization of alternative methods for measuring pollutant emissions, such as a predictive method. Machine Learning (ML) is a branch of Artificial Intelligence (AI) that has proven successful in various industries for prediction tasks. Therefore, the Nam Phong Power Plant is considering developing an ML-based NOx prediction model as a more cost-effective alternative to CEMS.

In recent years, the studies of predicting NOx emissions from gas turbines, diverse algorithms and methodologies have been explored across several studies. Kaya *et al.* (2019) introduced a publicly available dataset and employed Extreme Learning Machines (ELMs), emphasizing feature selection based on linear projection weights. Their work highlighted the importance of recognizing that the Machine Learning principle of "one size does not fit all" is applicable when employing decision fusion schemes. Another study by Rezazadeh (2020) delved into adaptive algorithms for NOx prediction, proposing the K-Nearest Neighbor (K-NN) algorithm and stressing the dynamic nature of power generation, advocating for accurate datasets. Kochueva and Nikolskii (2021) proposed a combined model using symbolic regression models and fuzzy classification, achieving metrics surpassing previous works. Huang *et al.* (2022) introduced a Neural Network model with an adjustable intermediate layer that demonstrated improved accuracy and adaptability, considering humidity effects. Finally, Chawathe (2021) underscored the importance of explainability in predictive

emission monitoring systems, achieving numerical accuracy comparable to less interpretable models.

Despite these advancements, certain research gaps persist. A notable limitation is the concentration of studies on specific types of gas turbines, indicating the need for broader applications across various power plant configurations (Chien, 2003). Additionally, the importance of ambient weather conditions as a key feature in emission prediction has been underscored in existing researches (Rezazadeh, 2020; Potts *et al*., 2023; Chawathe, 2021). Consequently, there is a compelling need to conduct additional studies within the specific context of the Nam Phong Power Plant, Thailand. This facility operates as a gas-fired combined-cycle power plant and employs steam injection systems for NOx control, a unique aspect not explored in previous studies. Investigating the applicability and effectiveness of NOx prediction with Machine Learning in such a specialized setting is crucial for a comprehensive understanding of its capabilities in this specific domain.

This study involves a thorough exploration of a diverse set of Machine Learning algorithms, encompassing six different approaches. The study employs Linear Regression, Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (K-NN), and Backpropagation Multilayer Perceptron Neural Network (BPMLP-NN) to predict NOx emissions. The performance of these ML algorithms is assessed using various statistical metrics, and the relationships between input variables and output predictions are scrutinized through feature importance analysis.

## MATERIALS AND METHODS

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) as shown in Figure 1, that enables systems to learn and improve from experience without explicit programming (Brown, 2021). Unlike traditional programming, which relies on step-by-step instructions, ML takes a data-driven approach. By analyzing patterns in data, ML systems can make informed decisions or predictions based on what they have learned. A fundamental concept in ML is the notion of a "model." A model serves as a representation of the patterns and relationships found within a given dataset. This process involves training the ML system using labeled data, where both input data and corresponding outcomes are provided. During training, the internal parameters of the system get adjusted to minimize any discrepancies between its predictions and actual outcomes. Once effectively trained, this model can then be utilized to make accurate predictions on new datasets that have not been previously encountered or analyzed.

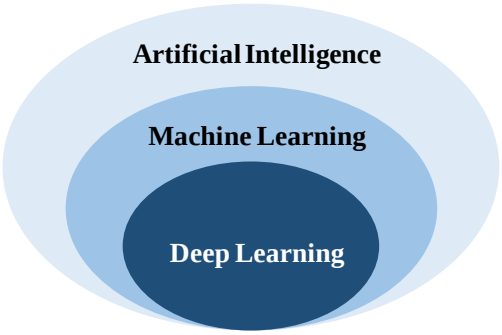


**Figure 1** Machine Learning a subfield of Artificial Intelligence

To accomplish the study's goals, the NOx emission prediction models for Nam Phong Power Plant are constructed through a multi-step methodology, depicted in Figure 2. This figure visually outlines the sequential process involved in developing the models, emphasizing the utilization of a ML approach for accurate predictions.

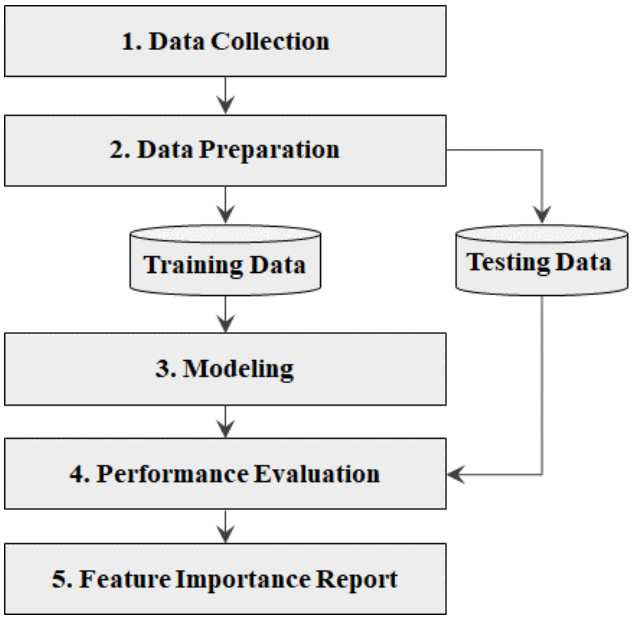


**Figure 2** Machine Learning development approach

## 1. Data Collection

The NOx prediction model was developed using input data obtained from gas turbine unit-21 (GT-21) and gas turbine unit-22 (GT-22) operational data during 1st to 30th September 2022. The data had a 1-minute sampling interval with 43,200 entries and 46 input parameters per unit, as listed in Table 1

**Table 1** Input parameters

| Item | Input Parameters | Unit | GT#21 | | | | GT#22 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | MEAN | STD. | MAX. | MIN. | MEAN | STD. | MAX. | MIN. |
| 1 | Gas Turbine Generation | MW | 79.3 | 4.9 | 99.8 | 60.1 | 79.2 | 4.9 | 102.3 | 60.3 |
| 2 | Gas Turbine Reactive Power | MVAR | 7.1 | 8.8 | 39.8 | -15.9 | 6.4 | 8.9 | 38.7 | -16.7 |
| 3 | Generator Current | KA | 3.3 | 0.2 | 4.3 | 2.5 | 3.3 | 0.2 | 4.3 | 2.5 |
| 4 | Generator Voltage | KV | 13.7 | 0.2 | 14.3 | 13.2 | 13.7 | 0.2 | 14.3 | 13.2 |
| 5 | Generator Power Factor | - | 0.5 | 0.9 | 1.0 | -1.0 | 0.5 | 0.9 | 1.0 | -1.0 |
| 6 | Fuel Gas Pressure | kg/cm2 | 23.1 | 0.1 | 23.9 | 22.6 | 23.1 | 0.1 | 23.8 | 22.6 |
| 7 | Fuel Command | % | 49.9 | 0.9 | 52.9 | 45.4 | 51.1 | 1.0 | 55.0 | 46.2 |
| 8 | Gas Turbine Speed | RPM | 2999.9 | 2.0 | 3005.6 | 2992.9 | 2999.9 | 2.0 | 3006.4 | 2992.9 |
| 9 | Inlet Guide Vane Command | % | 4.3 | 5.6 | 73.5 | 0.0 | 7.0 | 6.2 | 47.9 | 0.0 |
| 10 | Inlet Guide Vane Feedback | % | 4.2 | 5.6 | 73.3 | -0.3 | 7.2 | 6.1 | 48.9 | -0.9 |
| 11 | Exhaust Damper Command | % | 100.0 | 0.0 | 100.0 | 100.0 | 100.0 | 0.0 | 100.0 | 100.0 |
| 12 | Exhaust Damper Feedback | % | 103.0 | 0.1 | 103.3 | 102.6 | 103.8 | 0.0 | 104.6 | 103.0 |
| 13 | Compressor Inlet Air Temperature | ºC | 26.0 | 2.6 | 34.1 | 21.3 | 25.8 | 2.6 | 34.1 | 20.9 |
| 14 | Rotor Cooling Air Temp. (Right) | ºC | 186.7 | 4.8 | 205.8 | 173.7 | 192.6 | 5.8 | 330.4 | 179.1 |
| 15 | Rotor Cooling Air Temp. (Left) | ºC | 187.6 | 5.0 | 207.7 | 174.5 | 192.4 | 5.7 | 329.0 | 179.2 |
| 16 | Disc Cavity Row-2 Temp. (Right) | ºC | 359.4 | 5.0 | 371.3 | 336.5 | 333.4 | 4.4 | 348.0 | 313.1 |
| 17 | Disc Cavity Row-2 Temp. (Left) | ºC | 412.5 | 7.3 | 425.1 | 372.9 | 360.3 | 4.7 | 373.3 | 338.6 |
| 18 | Disc Cavity Row-3 Temp. (Right) | ºC | 325.8 | 6.0 | 340.2 | 297.3 | 365.5 | 6.8 | 378.5 | 326.2 |
| 19 | Disc Cavity Row-3 Temp. (Left) | ºC | 329.0 | 5.7 | 342.8 | 302.5 | 383.8 | 7.1 | 397.2 | 343.7 |
| 20 | Disc Cavity Row-4 Temp. (Right) | ºC | 381.1 | 9.7 | 397.0 | 333.1 | 329.3 | 7.2 | 361.4 | 291.6 |
| 21 | Disc Cavity Row-4 Temp. (Left) | ºC | 406.1 | 9.4 | 418.8 | 357.4 | 360.8 | 8.8 | 398.7 | 314.6 |
| 22 | Blade Path Temp. (Average) | ºC | 554.7 | 13.9 | 564.7 | 476.2 | 558.7 | 13.6 | 598.9 | 482.1 |
| 23 | Exhaust Gas Temp. (Average) | ºC | 542.8 | 10.7 | 546.6 | 482.0 | 542.5 | 11.5 | 548.4 | 478.1 |
| 24 | Fuel Gas Flow | kNm3/h | 28.2 | 1.2 | 32.6 | 23.5 | 28.1 | 1.1 | 33.5 | 23.5 |
| 25 | Fuel Gas Supply Temperature | ºC | 30.0 | 1.7 | 35.2 | 25.2 | 29.8 | 1.7 | 34.9 | 25.2 |
| 26 | Steam Injection Flow | T/H | 4.5 | 1.9 | 14.0 | 2.5 | 4.2 | 1.9 | 14.3 | 2.3 |
| 27 | Water Injection Flow | T/H | 0.0 | 0.1 | 3.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 28 | Compressor Outlet Air Temp. | ºC | 368.3 | 5.5 | 390.9 | 353.5 | 370.8 | 5.5 | 393.5 | 356.9 |

**Table 1** (Continuous)

| Item | Input Parameters | Unit | GT#21 | | | | GT#22 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | MEAN | STD. | MAX. | MIN. | MEAN | STD. | MAX. | MIN. |
| 29 | Combustor Shell Pressure | kg/cm2 | 9.4 | 0.3 | 11.6 | 8.7 | 9.5 | 0.3 | 11.4 | 8.9 |
| 30 | H2 Cooling Water Temp (Sensor-33) | oC | 45.0 | 0.4 | 47.1 | 43.8 | 42.6 | 2.1 | 48.2 | 37.4 |
| 31 | H2 Cooling Water Temp (Sensor-34) | oC | 44.3 | 0.2 | 45.3 | 43.4 | 42.8 | 2.0 | 48.0 | 37.6 |
| 32 | Steam Injection Temperature | oC | 445.0 | 10.9 | 456.3 | 408.8 | 442.4 | 10.1 | 451.7 | 404.6 |
| 33 | Compressor Index Pressure | mmH2O | 988.0 | 56.0 | 1634.0 | 943.2 | 1000.4 | 71.3 | 1357.9 | 882.6 |
| 34 | Raw NOx | PPM | 96.8 | 6.8 | 102.5 | 67.8 | 97.2 | 7.5 | 111.0 | 66.4 |
| 35 | Excessive Oxygen | % | 15.2 | 0.2 | 16.1 | 15.0 | 15.2 | 0.2 | 16.0 | 14.6 |
| 36 | Compressor Inlet Air Flow | kg/s | 323.0 | 7.7 | 403.3 | 314.1 | 324.9 | 10.9 | 373.4 | 305.6 |
| 37 | Fuel Gas Diff. Pressure | mmH2O | 2477.7 | 200.3 | 3296.0 | 1685.1 | 2468.0 | 200.5 | 3482.5 | 1684.2 |
| 38 | Inlet Air Diff. Pressure | mmH2O | 19.3 | 1.6 | 43.0 | 17.9 | 20.1 | 1.6 | 42.1 | 18.3 |
| 39 | NOx correction by 7% O2 | PPM | 235.5 | 13.6 | 250.6 | 173.9 | 234.9 | 15.2 | 268.2 | 171.2 |
| 40 | HRSG Inlet Gas Pressure | mmH2O | 95.0 | 65.0 | 201.6 | -3.1 | 200.5 | 11.2 | 275.2 | 186.0 |
| 41 | HRSG Outlet Gas Pressure | mmH2O | 9.6 | 0.7 | 15.5 | -3.8 | 9.9 | 0.8 | 17.1 | 7.1 |
| 42 | HRSG Inlet Gas Temperature | oC | 541.8 | 10.3 | 546.4 | 483.3 | 537.5 | 10.6 | 549.2 | 478.3 |
| 43 | HRSG Outlet Gas Temperature | oC | 112.4 | 0.8 | 119.5 | 110.8 | 107.1 | 0.9 | 112.6 | 106.1 |
| 44 | Steam Turbine Generation | MW | 94.7 | 1.7 | 96.5 | 85.5 | 94.7 | 1.7 | 96.5 | 85.5 |
| 45 | Ambient Temperature | oC | 27.5 | 2.6 | 35.4 | 23.2 | 27.5 | 2.6 | 35.4 | 23.2 |
| 46 | Relative Humidity | % | 85.9 | 9.4 | 97.1 | 50.0 | 85.9 | 9.4 | 97.1 | 50.0 |

## 2. Data Preparation

Data preparation steps included data cleaning, normalization, and partitioning, with cleaning to remove missing or invalid data, normalization to scale data to a common range, and partitioning to split data into training and testing sets for ML model training and evaluation. A 70:30 splitting ratio (Witten and Frank, 2005) was employed in this study.

## 3. Modeling

Modeling plays a crucial role in the research process, providing the necessary analytical tools to analyze large datasets and derive meaningful insights. In order to predict NOx emissions from gas turbine units, this study employed a diverse range of ML algorithms that offer distinct strengths and perspectives for analysis. The selected algorithms utilized in this study include Linear Regression, known for its simplicity and ability to interpret results; Decision Tree, which provides a visual and intuitive approach; Random Forest, an ensemble method built on Decision Trees that offers powerful modeling capabilities; XGBoost, an advanced Gradient Boosting framework recognized for its speed and efficiency; K-NN, a non-parametric method based on proximity principles; as well as BPMLP-NN which mimics the interconnected structure of biological neural networks and is particularly effective at handling complex patterns.

In order to achieve optimal performance, the algorithms used in this study underwent thorough training using preprocessed data from previous stages. However, training alone does not guarantee accurate predictions. To further enhance predictive accuracy, a careful hyper-parameter tuning process was implemented using the “GridSearchCV” methodology, which is a function that comes in Scikit-learn’s package (Great Learning Team, 2023). This optimization procedure systematically works through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. The specific hyperparameters selected after comprehensive tuning for each respective algorithm are presented in Table 2.

By employing this multifaceted modeling approach and conducting rigorous optimization, it is ensured that a comprehensive and accurate prediction of NOx emissions from the gas turbine units is achieved.

**Table 2** Hyper-Parameters after tuning

| Algorithms | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN |
|---|---|---|---|---|---|---|
| Hyper-Parameters | Default | - criterion : mse<br>- max_dept : 8<br>- max_leaf_nodes : 100<br>- min_samples_leaf : 100<br>- min_simple_split : 10 | Default | Default | n : 19 | - input layer : 30 neurals<br>- hidden layers<br>: 128 neurals<br>: Act. Func. : Relu<br>- output layer<br>: 1 neural<br>: Linear<br>- epochs : 100<br>- loss : MAE<br>- Optimizer : Adam |

## 4. Performance Evaluation

Performance evaluation serves as a critical checkpoint in the study, where the efficacy of the established ML models in forecasting NOx emissions is rigorously tested. The evaluation framework utilized in this study is built on seven robust performance metrics that paint a comprehensive picture of the model's performance, not only in terms of accuracy but also the efficiency and variability of predictions.

These vital metrics encompass:

**4.1 Mean Absolute Error (MAE)**: providing insights on the average magnitude of errors between the true and predicted values. A lower MAE indicates better performance.

$$\boldsymbol{MAE} = \frac{1}{N}\sum_i^N \left|Y_i - \hat{Y}_i\right| \quad (1)$$

**4.2 Mean Absolute Percentage Error (MAPE)**: offering a scale-independent indicator of prediction accuracy which represents on percentage (%). A lower MAE indicates better performance.

$$\boldsymbol{MAPE} = \frac{100}{N}\sum_i^N \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \quad (2)$$

**4.3 Coefficient of Determination ($R^2$)**: depicting how well the predicted outcomes match with the actual values. The $R^2$ value ranges from 0 to 1. A higher value suggests that a larger proportion of the variability in the dependent variable is explained by the model, indicating a better fit.

$$\boldsymbol{R^2} = 1 - \frac{\sum_i^N (Y_i - \hat{Y}_i)^2}{\sum_i^N (Y_i - \bar{Y}_i)^2} \quad (3)$$

**4.4 Standard Deviation (σ)**: revealing the dispersion of prediction errors. A lower standard deviation in prediction errors indicates more consistent model performance, while a higher standard deviation suggests greater variability in the accuracy of predictions.

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum_i^N (Y_i - \hat{Y}_i)^2}{N}} \quad (4)$$

**4.5 Maximum Error**: denoting the highest deviation observed.

**4.6 Training Time**: highlighting the duration taken to train the model.

**4.7 Prediction Time**: representing the prediction time of the model.

Where $Y_i$ , $\hat{Y}_i$ , $\bar{Y}_i$ and $N$ denote the NOx value measured from CEMS, the NOx value predicted by ML model, the averaged NOx value measured by CEMS, and the number of samples, respectively.

Moreover, this investigation will undertake a Wilcoxon Signed Rank Test (Rosner *et al.*, 2006) of MAPE at a 5% significance level. This non-parametric test compares the median of paired differences in two datasets. The test assigns signs (+ or -) to the differences based on their direction and tests if the signs are randomly distributed. A p-value less than 0.05 means the median difference is significantly different from zero. This indicates a significant difference in prediction accuracy between two models or methods. A positive (+) difference means the

first model or method has higher MAPE, which is worse. A negative (-) difference means the second model or method has lower MAPE, which is better. The test provides insights into the relative performance of the models or methods.

**5. Feature Importance Report**

Feature importance reports were generated to determine the relative significance of input features in a NOx emission prediction model, highlighting the critical factors for accurate predictions.

## RESULTS AND DISCUSSION

**1. Predictive Model Performances**

The assessment of ML models predicting NOx levels in gas turbines GT-21 and GT-22 is depicted in Figure 3 and Figure 4. These visuals showcase the strong alignment between predicted values and actual measurements from the CEMS. The figure incorporates six Machine Learning models-Linear Regression, Decision Tree, Random Forest, XGBoost, K-NN, and BPMLP-NN. The time series data covers NOx emissions from 1st to 30th September, with blue and orange marks representing actual and predicted values, respectively. The level of alignment signifies model accuracy. This visualization underscores the efficacy of ML in predicting NOx emissions.

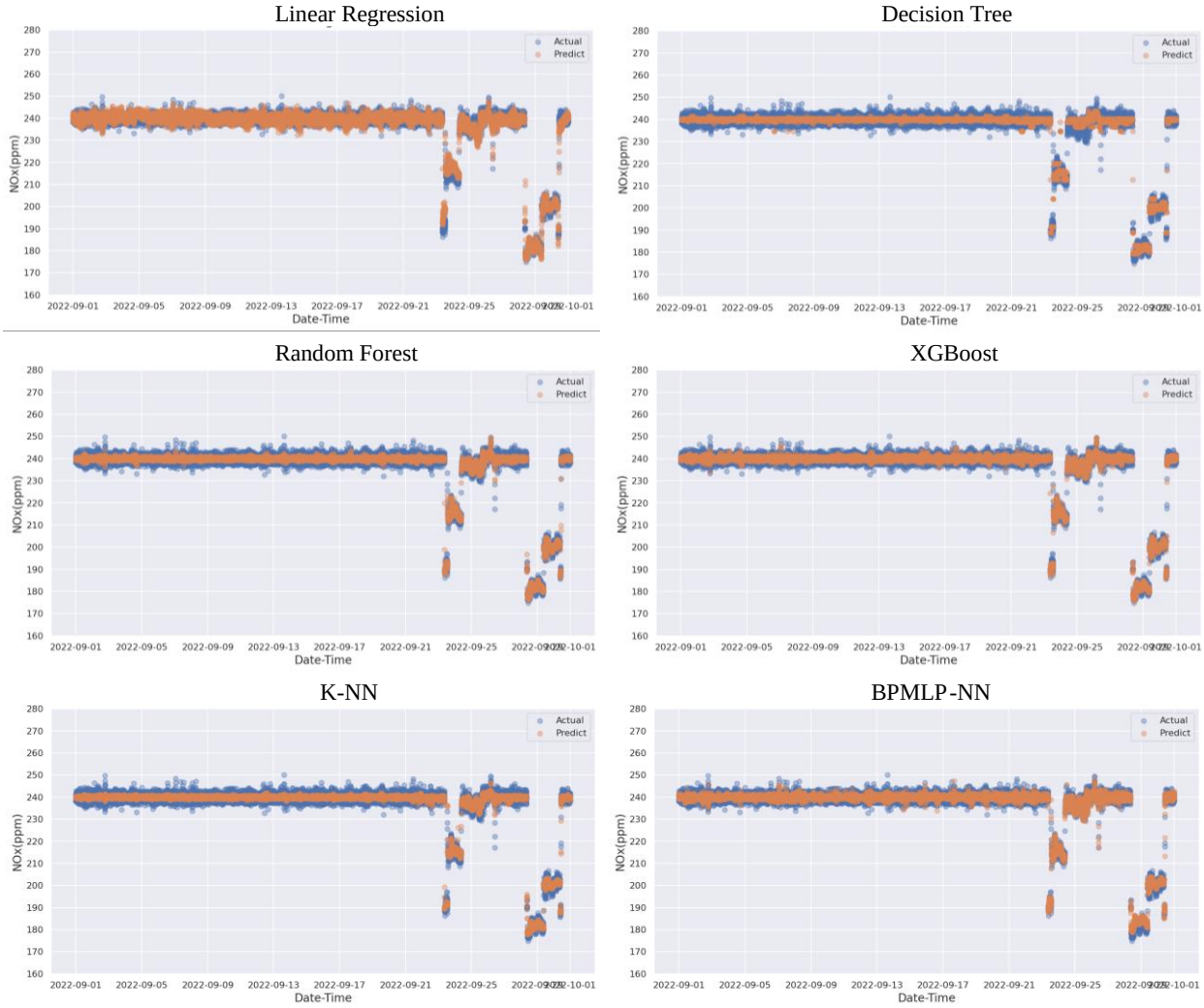


**Figure 3** NOx comparison between predicted values and actual measurement values of GT-21

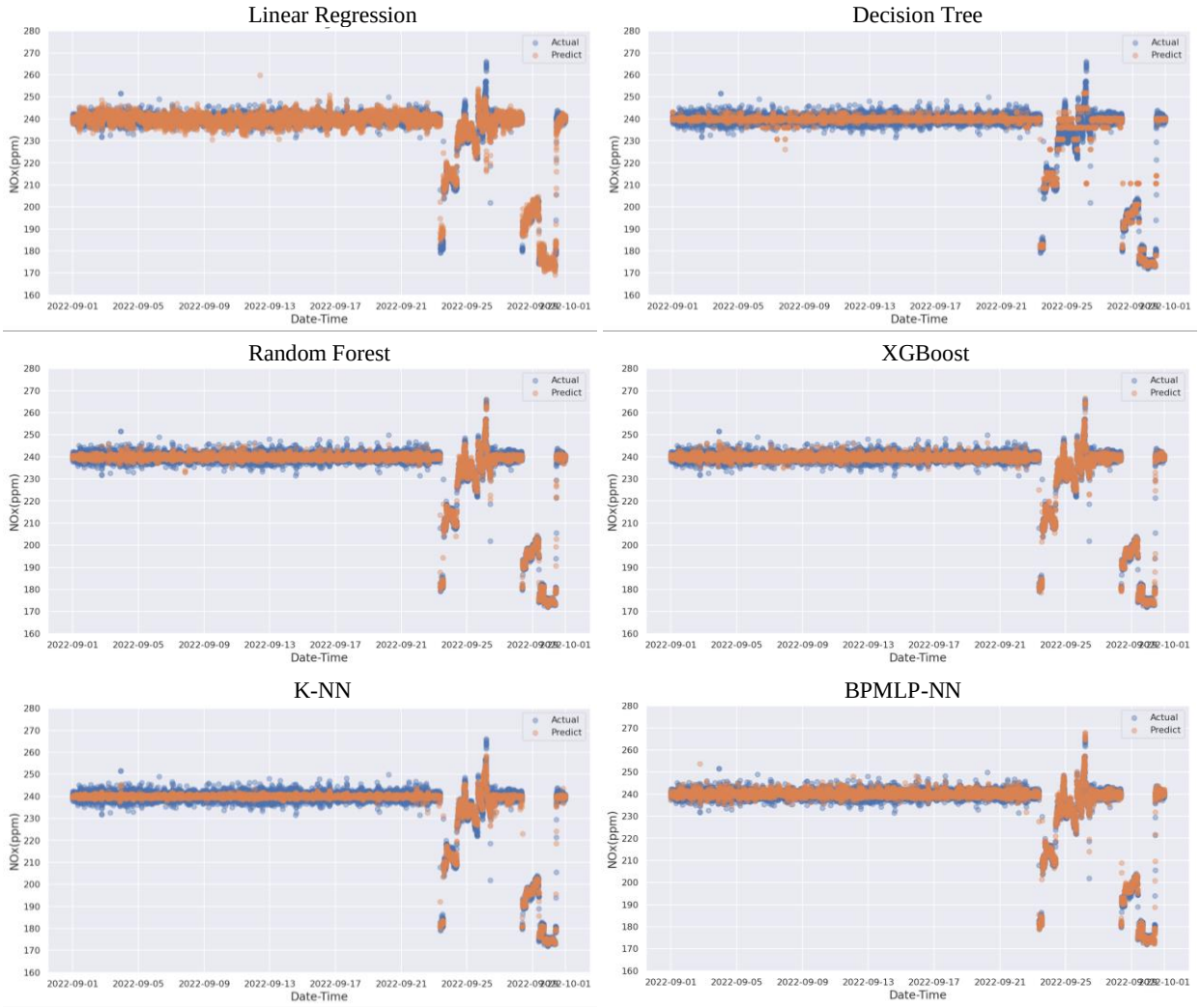


**Figure 4** NOx comparison between predicted values and actual measurement values of GT-22

### 1.1 Predictive Model Performance of GT-21

For GT-21, performance metrics in Table 3 highlight the superiority of the Random Forest model. It exhibits the lowest MAE at 1.184 ppm, emphasizing its high accuracy. Random Forest also excels in MAPE at 0.505 % and $R^2$ at 0.986, underscoring its predictive accuracy and data variance explanation. Additionally, it minimizes prediction errors with a low Standard Deviation of Error (ERR_STDev) of 1.557 ppm.

**Table 3** Performance Metrics of GT-21

| Parameters | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN |
|---|---|---|---|---|---|---|
| MAE (ppm) | 1.694 | 1.379 | **1.184** | 1.204 | 1.285 | 1.236 |
| MAPE (%) | 0.724 | 0.591 | **0.505** | 0.513 | 0.549 | 0.526 |
| $R^2$ | 0.973 | 0.977 | **0.986** | **0.986** | 0.984 | **0.986** |
| ERR_STDev (ppm) | 2.210 | 2.014 | **1.557** | 1.564 | 1.716 | 1.572 |
| ERR_Max (ppm) | 29.634 | 49.293 | 28.776 | **19.285** | 33.265 | 23.003 |
| ERR_Max (%) | 15.819 | 25.954 | 15.361 | **10.295** | 17.757 | 12.279 |
| Training_Time (sec) | 0.152 | 1.009 | 108.258 | 7.205 | **0.003** | 135.383 |
| Predict_Time (sec) | 0.005 | **0.003** | 0.451 | 0.017 | 42.534 | 0.372 |

Results of Wilcoxon Signed Rank tests are presented in Table 4 to investigate the significant difference in performance of algorithms of GT-21. The result further corroborates the superiority of the Random Forest model at a 5% significance level. The test compares the MAPE of each pair of models and assigns a sign (+ or -) based on which model has lower MAPE. A (+) sign means that the model in the row has significantly lower MAPE than the model in the column, indicating better prediction performance. A (-) sign means that there is no significant difference between the MAPE of the two models. The Random Forest model has all (+) signs, indicating that it has significantly lower MAPE than other models.

**Table 4** Wilcoxon Signed Rank Test of MAPE results of GT-21 at 5% significant level

| Algorithms | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN |
|---|---|---|---|---|---|---|
| Linear Regression | N/A | (-) | (-) | (-) | (-) | (-) |
| Decision Tree | (+) | N/A | (-) | (-) | (-) | (-) |
| **Random Forest** | **(+)** | **(+)** | N/A | **(+)** | **(+)** | **(+)** |
| XGBoost | (+) | (+) | (-) | N/A | (+) | (+) |
| K-NN | (+) | (+) | (-) | (-) | N/A | (-) |
| BPMLP-NN | (+) | (+) | (-) | (-) | (+) | N/A |

### 1.2 Predictive Model Performance of GT-22

For GT-22, the performance metrics are represented in Table 5. The Random Forest remains the top performer, with the lowest MAE (0.99 ppm.), MAPE (0.42%), and strong $R^2$ values at 0.992. It also excels in minimizing prediction errors with a low ERR_STDev at 1.374 ppm. Moreover, it showcases the lowest ERR_Max in this context, signifying robustness in handling extreme prediction errors.

**Table 5** Performance Metrics of GT-22

| Parameters | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN |
|---|---|---|---|---|---|---|
| MAE(ppm) | 1.925 | 1.365 | **0.990** | 1.038 | 1.134 | 1.051 |
| MAPE(%) | 0.823 | 0.584 | **0.420** | 0.441 | 0.482 | 0.453 |
| $R^2$ | 0.968 | 0.971 | **0.992** | 0.991 | 0.989 | 0.991 |
| ERR_STDev(ppm) | 2.733 | 2.571 | **1.374** | 1.418 | 1.615 | 1.397 |
| ERR_Max(ppm) | 49.875 | 55.480 | **17.143** | 20.945 | 36.409 | 22.728 |
| ERR_Max(%) | 18.744 | 20.850 | **8.494** | 10.378 | 19.507 | 12.177 |
| Training_Time(sec) | 0.026 | 0.636 | 141.265 | 9.432 | **0.006** | 128.638 |
| Predict_Time(sec) | 0.008 | **0.003** | 0.955 | 0.021 | 24.400 | 0.304 |

In alignment with the findings for GT-21, the Wilcoxon Signed Rank Test outcomes for GT-22, as presented in Table 6, further substantiate the supremacy of the Random Forest model at a 5% significance level. The Random Forest model displays exclusively (+) signs, underscoring its consistent and statistically significant superiority in achieving lower MAPE compared to alternative models.

**Table 6** Wilcoxon Signed Rank Test of MAPE results of GT-22 at 5% significant level

| Algorithms | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN |
|---|---|---|---|---|---|---|
| Linear Regression | N/A | (-) | (-) | (-) | (-) | (-) |
| Decision Tree | (+) | N/A | (-) | (-) | (-) | (-) |
| **Random Forest** | **(+)** | **(+)** | N/A | **(+)** | **(+)** | **(+)** |
| XGBoost | (+) | (+) | (-) | N/A | (+) | (+) |
| K-NN | (+) | (+) | (-) | (-) | N/A | (-) |
| BPMLP-NN | (+) | (+) | (-) | (-) | (+) | N/A |

In conclusion, Random Forest consistently stands out as the top model for predictive accuracy in both GT-21 and GT-22. However, this heightened accuracy comes at the cost of longer training times. Conversely, Decision Tree and XGBoost offer a balanced compromise between accuracy and speed, making them suitable choices when real-time predictions are essential. The choice of model should be made thoughtfully, aligning with the specific application requirements while considering both performance and time constraints.

## 2. Feature Importance Analysis

The feature importance for gas turbines GT-21 and GT-22 models is displayed in Table 7 and Table 8, where they are ranked by their top ten average scores. These scores, ranging from 0 to 100%, signify the degree of importance, with higher scores reflecting greater significance.

### 2.1 Feature Importance Analysis of GT-21

In the case of GT-21, different Machine Learning algorithms produce varying importance rankings for features. Linear Regression identifies "Steam Injection Flow" as the most important, while Decision Tree places great emphasis on "Steam Injection Temperature." Random Forest and XGBoost concur with the significance of "Steam Injection Temperature." In contrast, K-NN deviates, prioritizing "Steam Injection Flow." BPMLP-NN aligns with Decision Tree, Random Forest and XGBoost models in highlighting "Steam Injection Temperature." Across these algorithms, "Steam Injection Temperature," "Steam Injection Flow," and "Relative Humidity" consistently emerge as the top three features with the highest average importance scores.

**Table 7** Top 10 Feature importance of GT-21

| Input Parameters | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN | Average |
|---|---|---|---|---|---|---|---|
| **Steam Injection Temperature** | 5.334 | **92.844** | **88.477** | **80.879** | **27.384** | **18.375** | **52.216** |
| **Steam Injection Flow** | **15.459** | **3.283** | **3.596** | 3.214 | **61.727** | **16.931** | **17.368** |
| **Relative Humidity** | **13.077** | 0.096 | 0.183 | 0.186 | 0.635 | **14.979** | **4.859** |
| Ambient Temperature | 11.471 | 0.001 | 0.119 | 0.147 | 0.404 | 12.827 | 4.162 |
| Combustor Shell Pressure | **12.193** | 0.009 | 0.072 | 0.077 | 0.046 | 3.310 | 2.618 |
| Compressor Inlet Air Temp. | 1.068 | **3.319** | 2.588 | **5.578** | 0.293 | 2.385 | 2.539 |
| Steam Turbine Generation | 0.099 | 0.000 | **2.638** | **7.994** | 1.033 | 1.357 | 2.187 |
| Disc Cavity Row-3 Temp. | 8.815 | 0.000 | 0.679 | 0.252 | 0.185 | 1.171 | 1.850 |
| Disc Cavity Row-2 Temp. | 6.484 | 0.254 | 0.183 | 0.316 | 0.148 | 1.502 | 1.481 |
| Compressor Outlet Air Temp. | 5.548 | 0.000 | 0.141 | 0.061 | 0.214 | 1.517 | 1.247 |

### 2.2 Feature Importance Analysis of GT-22

For GT-22, the feature importance analysis again reveals variations depending on the algorithm used. Linear Regression points to "HRSG Inlet Gas Temperature" as the most significant feature, while Decision Tree designates "Steam Turbine Generation" with the highest importance. Random Forest concurs by emphasizing "Steam Turbine Generation," and XGBoost follows suit. In contrast, K-NN underscores "Steam Injection Flow," and BPMLP-NN assigns importance to "Relative Humidity." The top three features with the highest average importance scores for GT-22 are "Steam Turbine Generation," "Steam Injection Flow," and "Steam Injection Temperature."

**Table 8** Top 10 Feature importance of GT-22

| Input Parameters | Linear Regression | Decision Tree | Random Forest | XGBoost | K-NN | BPMLP-NN | Average |
|---|---|---|---|---|---|---|---|
| **Steam Turbine Generation** | 0.560 | **85.295** | **59.298** | **91.772** | 1.178 | 2.527 | **40.105** |
| **Steam Injection Flow** | 8.829 | **4.181** | **4.865** | **1.574** | **59.290** | **12.372** | **15.185** |
| **Steam Injection Temperature** | 4.803 | **5.757** | **29.728** | **3.823** | **24.434** | **13.289** | **13.639** |
| Relative Humidity | 8.160 | 3.493 | 3.108 | 1.474 | 1.403 | **14.961** | 5.433 |
| Blade Path Temp. | **12.746** | 0.055 | 0.106 | 0.049 | 0.167 | 6.944 | 3.345 |
| HRSG Inlet Gas Temp. | **14.506** | 0.003 | 0.069 | 0.045 | 0.155 | 2.006 | 2.797 |
| Ambient Temperature | 7.141 | 0.099 | 0.161 | 0.074 | 0.472 | 6.486 | 2.405 |
| Combustor Shell Pressure | **12.584** | 0.268 | 0.037 | 0.016 | 0.050 | 0.737 | 2.282 |
| Exhaust Gas Temp. | 4.787 | 0.012 | 0.088 | 0.025 | 0.094 | 8.266 | 2.212 |
| Fuel Gas Flow | 6.094 | 0.000 | 0.041 | 0.015 | 0.010 | 2.151 | 1.385 |

The consistency in the prominence of "Steam Injection Temperature" and "Steam Injection Flow" across different algorithms for both GT-21 and GT-22 underscores their crucial roles in predicting NOx emissions. However, the specific rankings may vary, highlighting the need to tailor feature selection to the chosen model. Furthermore, the interdependencies between variables suggest that simplifying the problem by eliminating certain dependent variables in future studies may enhance model accuracy. These findings provide valuable guidance for understanding the factors influencing NOx emissions in gas turbines and offer insights for refining predictive models.

## CONCLUSION

In conclusion, this research emphasizes the potential of Machine Learning, particularly Random Forest, for accurate forecasting of NOx emissions at the Nam Phong Power Plant. It aligns with Thailand's evolving environmental regulations and underscores the relevance of data-driven insights in emissions monitoring and mitigation efforts. While Random Forest excels in predictive accuracy, models like Decision Tree and XGBoost strike a balance between accuracy and speed, making them suitable for real-time predictions. The synergy of technological advancements and regulations highlights the significance of swiftly adopting NOx prediction systems at the Nam Phong Power Plant.

Moreover, the analysis of feature importance highlights the potential for reducing NOx emissions without compromising plant efficiency by controlling factors like steam injection temperature. These insights hold practical implications, benefiting public health, air quality, and minimizing the ecological impact, while also contributing to a more environmentally aware and sustainable future.

However, it’s important to note that the analysis used data from a limited time frame (Sep 1-30, 2022), raising concerns about long-term model performance representativeness. Future studies should include data from various seasons and years for extended model stability evaluation.

## ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the Electricity Generating Authority

of Thailand (EGAT) for the financial support of this study. Special appreciation goes to Mr. Somkate Thongthom, Chief of Maintenance, and the Nam Phong Power Plant team for their unwavering commitment and data accessibility, enriching the depth and credibility of this study.

## REFERENCES

Brown, S. 2021. **Machine learning, explained - MIT Sloan.** Artificial Intelligence. Available Source: https:// mitsloan.mit. edu/ ideas- made- to- matter/ machine-learning-explained, September 3, 2023.

Chawathe, S.S. 2021. Explainable predictions of industrial emissions, pp. 1-7. *In* **2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).** IEEE, New Jersey.

Chien, T.W., Chu, H., Hsu, W.C., Tseng, T.K., Hsu, C.H. and Chen, K.Y. 2003. A feasibility study on the predictive emission monitoring system applied to the Hsinta power plant of Taiwan power company. **Journal of the Air & Waste Management Association** 53(8): 1022-1028.

Department of Industrial Works. 2007. **Notification of Department of Industrial Works 2007 Title: Data Transmission into the system of Continuous Emission Monitoring System**. Government Gazette vol.124, Special Part 196. (dated October 10, 2007). (in Thai)

Great Learning Team. 2023. **Hyperparameter Tuning with GridSearchCV.** AI and Machine Learning. Available Source: https: / / www. mygreatlearning. com/ blog/gridsearchcv, September 5, 2023.

Huang, D., Tang, S., Zhou, D. and Hao, J. 2022. NOx emission estimation in gas turbines via interpretable neural network observer with adjustable intermediate layer considering ambient and boundary conditions. **Measurement** 189: 110429.

Kaya, H., Tüfekci, P. and 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.

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

Ministry of Industry. 2022. **Notification of Department of Industrial Works 2022 Title: Mandating factories to install special tools or equipment for reporting air pollutants emitted from factory stacks.** Government Gazette vol. 139, Special Part 131. (dated April 1, 2022). (in Thai)

Potts, R., Hackney, R. and Leontidis, G. 2023. Tabular machine learning methods for predicting gas turbine emissions. **Machine Learning & Knowledge Extraction** 5(3): 1055-1075.

Rezazadeh, A. 2020. Environmental pollution prediction of NOx by predictive modelling and process analysis in natural gas turbine power plants. **ArXiv** 1: 8978.

Rosner, B., Glynn, R.J. and Lee, M.L.T. 2006. The Wilcoxon signed rank test for paired comparisons of clustered data. **Biometrics** 62:185-192.

United States Environmental Protection Agency. 2023. **Basic information about NO2.** Nitrogen Dioxide (NO2) Pollution. Available Source: https: / / www. epa. gov / no2-pollution/basic-information-about-no2, September 4, 2023.

Witten, I. H. and Frank, E. 2005. **Data Mining: Practical Machine Learning Tools and Techniques.** 2nd ed. Elsevier, Amsterdam.