Original Research Article: full paper

(2024), «EUREKA: Physics and Engineering»
Number 1

PREDICTION OF AIR POLLUTION FROM POWER GENERATION USING MACHINE LEARNING

Thongchai Photsathian
Division of Instrumentation and Automation Engineering Technology
Faculty of Engineering and Technology[1]

Thitiporn Suttikul
Division of Chemical Process Engineering Technology
Faculty of Engineering and Technology[1]
The Plasma and Automatic Electric Technology Research Group[1]

Worapong Tangsrirat✉
Department of Instrumentation and Control Engineering
School of Engineering
King Mongkut's Institute of Technology Ladkrabang (KMITL)
Chalongkrung Road, Ladkrabang district, Bangkok, Thailand, 10520
worapong.ta@kmitl.ac.th

[1]*King Mongkut's University of Technology North Bangkok*
3371 National Highway, Ban Khai district, Rayong, Thailand, 21120

✉**Corresponding author**

Abstract

Electrical energy is now widely recognized as an essential part of life for humans, as it powers many daily amenities and devices that people cannot function without. Examples of these include traffic signals, medical equipment in hospitals, electrical appliances used in homes and offices, and public transportation. The process that generates electricity can pollute the air. Even though natural gas used in power plants is derived from fossil fuels, it can nevertheless produce air pollutants involving particulate matter (PM), nitrogen oxides (NO_x), and carbon monoxide (CO), which affect human health and cause environmental problems. Numerous researchers have devoted significant efforts to developing methods that not only facilitate the monitoring of current air quality but also possess the capability to predict the impacts of this increasing rise. The primary cause of air pollution issues associated with electricity generation is the combustion of fossil fuels. The objective of this study was to create three multiple linear regression models using artificial intelligence (AI) technology and data collected from sensors positioned around the energy generator. The objective was to precisely predict the amount of air pollution that electricity generation would produce. The highly accurate forecasted data proved valuable in determining operational parameters that resulted in minimal air pollution emissions. The predicted values were accurate with the mean squared error (MSE) of 0.008, the mean absolute error (MAE) of 0.071, and the mean absolute percentage error (MAPE) of 0.006 for the turbine energy yield (TEY). For the CO, the MSE was 2.029, the MAE was 0.791, and the MAPE was 0.934. For the NO_x, the MSE was 69.479, the MAE was 6.148, and the MAPE was 0.096. The results demonstrate that the models developed have a high level of accuracy in identifying operational conditions that result in minimal air pollution emissions, with the exception of NO_x. The accuracy of the NO_x model is relatively lower, but it may still be used to estimate the pattern of NO_x emissions.

Keywords: air pollution, artificial intelligence, prediction, power plant, machine learning (ML).

DOI: 10.21303/2461-4262.2024.003148

1. Introduction

The environment suffered from a double-edged effect as a result of rising energy demand: increased combustion of fossil fuels and biomass and increased emissions of carbon and flue gases [1, 2]. The majority of countries have made an effort to lower air pollution emissions due to concerns about climate change and global warming [2]. Nitrogen oxides (NO_x), comprising nitrogen oxide (NO_2), nitric oxide (NO), and carbon monoxide (CO), are considered the primary atmospheric pollutants due to they can be harmful to humans when inhaled in high concentrations [3], and cause environmental problems such as acid rain, photochemical smog, tropospheric ozone, and ultimately global warming [4]. The combustion process in many different industries, such as gas turbines and boilers in power plants, is a key contributor to the hazardous pollutants (NO_x, CO, and PM, etc.) emitted into the environment [5, 6].

The objective of this work is to forecast air pollution released from a power plant using artificial intelligence (AI). Research on the use of AI technology to forecast air pollution using various data sets has been published in the literature. A systematic review and a bibliographic perspective on air pollution detection are detailed in [7, 8]. Additionally, machine learning algorithms-based air pollution detection and forecasting methods have been suggested in [9, 10]. In this study, let's primarily concentrate on forecasting air pollution from a gas turbine power plant that utilized natural gas as fuel. The amount of air pollution, such as NO_x and CO, and the efficiency of producing electricity, such as turbine energy yield (TEY) was forecast based on a database from [11].

2. Materials and methods

2. 1. Experimental procedures

Fig. 1 shows the flow chart that provides the details of the experimental procedures utilized in this study. After preprocessing the initial dataset, all of the data were split into two sets in accordance with an 80:20 division ratio. A training set is represented by an 80 % data ratio, while a test set is denoted by a 20 % data ratio. The training set was subsequently applied to create a model (model creation). In order to evaluate the performance of the model (model evaluation), the test set was utilized to put it to the test. In the event that the testing results were deemed unsatisfactory, a new model was created. Conversely, the model was retained and utilized for future model development if the results were found to be satisfactory.

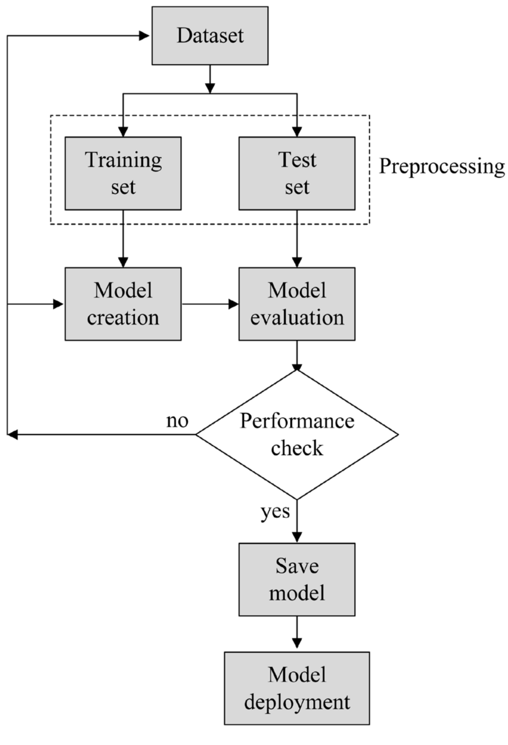

Fig. 1. Flow chart of the experimental procedures

Detailed descriptions of each procedure are provided as follows.

2. 2. Explanation of gas turbine power plant

A gas turbine power plant is classified as a turbine-based power plant since it generates electricity. By combusting fossil fuel with air, high-temperature and high-pressure gas was produced, which is used to power a turbine. The operation of a gas turbine power plant is shown in **Fig. 2**.

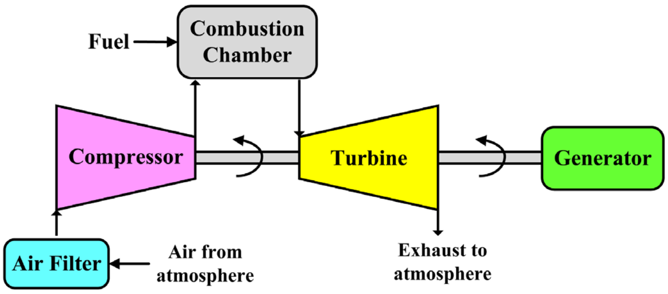

Fig. 2. Operation of a gas turbine power plant

To remove various impurities attached to the air, the process begins with air passing through an air filter powered by an air compressor. Subsequently, an air compressor presses the filtered air, raising its temperature and pressure. The compressed air and fossil fuel are combined and burned in a combustion chamber to produce high-pressure high-temperature gas. When the hot gas flows through the gas turbine, it causes the turbine to rotate, and a shaft of this turbine is connected to a generator to produce electricity. The air compressor is rotated by another section of the shaft. The hot gas that flows through the gas turbine has lower pressure and temperature, and then is released into the atmosphere.

2. 3. Data preparation

Before entering model creation, preliminary data is processed by checking missing values and outlines. The data to be utilized is comprised of information collected by sensors which are positioned at eleven different locations throughout the generator, as depicted in **Fig. 3**. The data has been classified into three categories shown in **Table 1**. The data concerning process and environment parameters are categorized as input or initial variables, while the data pertaining to prediction parameters is categorized as a dependent variable. There are 36,733 data points in total for each type. Horizontal data presents the minimum, mean, maximum, unit, and total amount of data, separated by data type, and includes the following details.

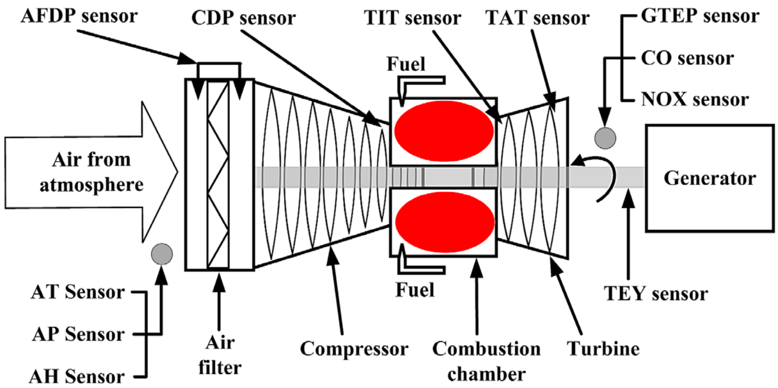

Fig. 3. Position of sensors on the generator

Table 1
Data parameters

–	Environment parameters			Process parameters					Prediction parameters		
	AT	AP	AH	AFDP	GTEP	TIT	TAT	CDP	TEY	CO	NO_x
Count	36733	36733	36733	36733	36733	36733	36733	36733	36733	36733	36733
Min	6.2348	985.85	24.085	2.0874	17.698	1000.80	511.04	1.0887	15.618	2.2626	11.678
Mean	17.712	1013.07	77.867	3.9255	25.563	1081.428	546.158	12.06	133.51	2.3724	65.293
Max	37.103	1036.60	100.20	7.6106	40.716	1100.90	550.61	15.159	179.50	44.103	119.91
Unit	°C	mbar	%	mbar	mbar	°C	°C	mbar	MWH	mg/m^3	mg/m^3

2. 3. 1. Data on environment parameters (input parameters or initial variables)

The data consist of the following parameters:

– Ambient temperature (AT) is the ambient operating temperature of a gas turbine (°C).

– Ambient pressure (AP) is the operating pressure of a gas turbine (mbar).

– Ambient humidity (AH) is the relative humidity in the air of a gas turbine (%).

2. 3. 2. Data on process parameters (input parameters or initial variables)

These parameters consist of the following variables:

– Air filter difference pressure (AFDP) is the different in air between entering and exiting the filter (mbar).

– Gas turbine exhaust pressure (GTEP) is the pressure of exhaust gas that exits the turbine into atmosphere (mbar).

– Turbine inlet temperature (TIT) is the temperature of the gas that exits the combustion chamber before entering the turbine (°C).

– Turbine after temperature (TAT) is the temperature of the gas that exits the combustion chamber and the turbine (°C).

– Compressor discharge pressure (CDP) is an air pressure at the on/off valve used to reduce the amount and pressure of air that exits the compressor before entering the combustion chamber (mbar).

2. 3. 3. Data on prediction parameters (prediction parameters or dependent variables)

The following parameters are incorporated into the data:

– Turbine energy yield (TEY) is an energy yield obtained from a turbine (MWH).

– Carbon monoxide (CO) is the amount of carbon monoxide produced by the incomplete combustion of fossil fuels (mg/m^3).

– Nitrogen oxide (NO_x) is the amount of nitrogen oxide produced by combustion, which consists of NO_2 and NO (mg/m^3).

2. 3. 4. Dimensionality reduction of data

Dimensionality reduction is the process of lowering the amount of input data by converting or deleting the data that is irrelevant or less significant. It does not reduce the amount of data in the input samples, but rather the number of columns or features of the data. Consequently, processing time is reduced, and fewer resources are used.

This study focused on data correlation and dimensionality reduction to improve forecasting performance. First, the initial variables in the columns are sorted according to their high correlation with the target or dependent variables that were utilized for prediction. Nonetheless, in the case of a low correlation, the variable will be removed. In other words, if the correlation value is nearly zero, it almost certainly indicates that there is no correlation at all, and the variable can be deemed eliminated. Conversely, a correlation value that is close to one suggests a high level of correlation.

Fig. 4 shows the correlation between the initial and dependent variables. With a correlation value of 0.015, the TEY and AT exhibit low correlation. Thus, the AT can be eliminated when predicting the TEY. On the contrary, the correlation between TEY and CDP is 0.99, indicating a high level of correlation. As a result, it ought to forecast the TEY using the CDP. (1) was used to calculate the correlation.

Table 2 provides a rule of thumb for interpreting the size of a correlation [12], which corresponds to the correlation determined using (1).

After performing dimensionality reduction and correlation assessment, the prepared data was divided into two groups, with 80 % of the data being used as a training set and 20 % as a test set, in a ratio of 80:20:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \tag{1}$$

where r is the correlation coefficient, x_i is a sample of x-variable, $\bar{x}$ is the mean of x-variable, y_i is a sample of y-variable, and $\bar{y}$ is the mean of y-variable. In (1), the variable y can be defined as:

$$y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n, \tag{2}$$

where y is the predicted value of the target or a dependent variable, w_k (k = 1, 2, …, n) is the numeric coefficient, or the weight of the model, and x_k is the predictor value or the independent variable.

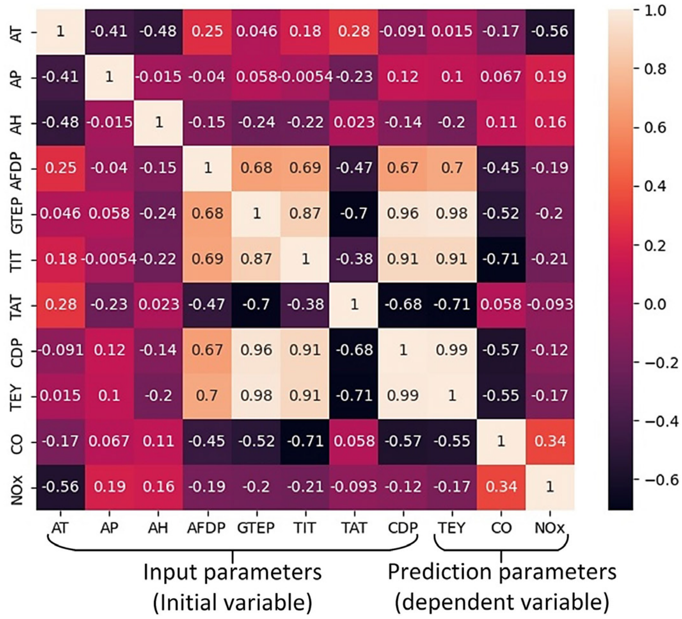

Fig. 4. Correlation between the initial and dependent variables

Table 2

Rule of thumb for interpreting the size of a correlation [12]

Correlation size	Interpretation
0.90 to 1.00 (–0.90 to –1.00)	Very high positive (negative) correlation
0.70 to 0.90 (–0.70 to –0.90)	High positive (negative) correlation
0.50 to 0.70 (–0.50 to –0.70)	Moderate positive (negative) correlation
0.30 to 0.50 (–0.30 to –0.50)	Low positive (negative) correlation
0.00 to 0.30 (–0.00 to –0.30)	Negligible correlation

2. 4. Model creation

This step involves creating a model using all of the data from the previous data preparation and then training it with multiple linear regression (MLR) to create a model for forecasting [13]. MLR is a linear regression model with multiple independent variables using Equation (2), whereas simple linear regression models linear dependency between a target or dependent variable and an initial or independent variable [14, 15].

2. 5. Model evaluation

This step evaluates the performance of the model in terms of accuracy using the test set data. The following is the method used to assess the performance of the model.

2. 5. 1. Mean squared error (MSE)

The mean squared error (MSE) is the average squared difference between the actual and predicted values using the following equation:

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_i^*\right)^2, \tag{3}$$

where n is the number of observations, y_i is the actual value of the i-th observation, and y_i^* is the predicted value of the i-th observation.

2. 5. 2. Mean absolute error

The mean absolute error (MAE) is the average absolute difference discrepancy between the actual values of the data and the predicted values of the model, according to the following equations:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-y_i^*\right|. \tag{4}$$

2. 5. 3. Mean absolute percentage error

The mean absolute percentage error (MAPE) is the average absolute percent difference between the actual and predicted values based on the following relationship:

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i^*}{y_i}\right|. \tag{5}$$

Once the performance of the model has been assessed to verify that it satisfies the desired accuracy, it can be employed for subsequent predictive purposes. However, in the event that the model's performance is evaluated as lacking in accuracy, it will be enhanced through data preparation or the generation of a new model until its accuracy is deemed to be satisfactory.

3. Results and discussions

The experiments conducted in this study used the Python programming language in conjunction with Google Colab and the Scikit-learn library to create the model. **Table 1** lists all data parameters utilized in this investigation; of these, only three –TEY, CO, and NO_x – are pertinent to prediction. As shown in **Fig. 4**, the correlation coefficients between the TEY and the CDP (0.99), GTEP (0.98), and TIT (0.91) are extremely high values. The range of 0.91 to 1.00 for the three correlation values suggests an exceptionally high positive (negative) correlation. In order to reduce computation time and computer burden, the TEY was removed from the initial variable as its correlation with the AT (0.015) was a remarkably low. Therefore, in this research, the input variables that exhibited a significantly low correlation with the prediction parameters were eliminated as follows:

1. The correlation value of 0.015 led to the elimination of the AT, and the TEY prediction was made using the AP, AH, AFDP, GTEP, TIT, TAT, CDP, CO, and NO_x.

2. The correlation values of 0.058 and 0.067 led to the elimination of the TAT and AP, and the CO prediction was made using the AT, AH, AFDP, GTEP, TIT, CDP, NO_x, and TEY.

3. The correlation value of -0.093 led to the elimination of the TAT, and the NO_x prediction was made using the AT, AP, AH, AFDP, GTEP, TIT, CDP, TEY, and CO.

After that, a total of 36,733 data points were divided into two sets with a ratio of 80:20. The test set comprised 7,347 data points, which accounted for 20 % of the total. The training set comprised 29,386 data points, or the remaining 80 %. As shown in **Table 3**, the MLR was utilized to create the model using the training set of data, and the test set of data was used to evaluate its performance in terms of MSE, MAE, and MAPE. **Table 4** presents a comparison between of the actual and predicted values of the TEY, CO, and NO_x with 10 times of prediction examples. The funding revealed that the actual and predicted values were relatively similar. Among the models, the TEY exhibited the maximum level of accuracy, as evidenced by its MSE value of 0.008, MAE value of 0.071, and MAPE value of 0.006. A reasonable level of accuracy was demonstrated by the CO and NO_x prediction performance, with MSE values of 2.029 and 69.479, MAE values of 0.791 and 6.148, and MAPE values of 0.934 and 0.096, respectively. The results indicate that the NO_x exhibited a relatively low level of accuracy, necessitating additional refinement of the model to enhance its accuracy. Nevertheless, it is possible to approximate the pattern of NO_x emissions produced by electricity generation. The study's limitation is that it employed data acquired from a website and then collected further data to improve the accuracy of the model.

Table 3
Comparison of the accuracy of the TEY, CO, and NO_x prediction performances

Parameter	Model evaluation		
	MSE	**MAE**	**MAPE**
TEY	0.008	0.071	0.006
CO	2.029	0.791	0.934
NO_x	69.479	6.148	0.096

Table 4
Comparisons of the actual and predicted values of TEY, CO, NO_x

No.	TEY (MWH)		CO (mg/m^3)		NO_x (mg/m^3)	
	actual	**predicted**	**actual**	**predicted**	**actual**	**predicted**
1	11.721	11.704	1.810	1.997	48.876	52.570
2	11.155	11.171	1.941	2.989	68.584	69.239
3	13.433	13.469	1.365	1.179	55.400	66.620
4	13.274	13.424	1.321	0.629	52.136	55.410
5	11.505	11.528	3.455	2.444	57.123	56.447
6	11.260	11.204	4.232	3.844	69.161	73.034
7	10.695	10.623	8.418	6.497	80.834	80.599
8	12.481	12.491	0.289	0.855	66.639	55.760
9	11.810	11.901	1.231	1.830	63.983	69.065
10	13.913	13.963	0.574	1.085	67.361	68.152

In future work, the multiple linear regression employed in this study will be applied to improve the performance of deep neural networks (DNNs) in predicting TEY, CO, and NO_x emissions from fossil fuel-burning gas turbine power plants. With the collaboration of experts and the devolved models, power plant planning and management can be facilitated, leading to more efficient electricity production and lower air pollution emissions. The accurate anticipated data was helpful in estimating operational parameters that led to low air pollution emissions.

4. Conclusions

In this study, a model for CO and NO_x emission prediction was developed using multiple linear regression. The prediction factors involved not only the amounts of air pollutants (e.g., CO and NO_x) produced by burning fossil fuels in a gas turbine power plant, but also the efficiency of producing electricity, which includes the TEY. Prior to the model creation, the initial data was prepared using correlation and dimensionality reduction. The results demonstrate that processing time and data size were reduced when the initial variables with low correction were diminished. Following the reduction in dimensionality, the nine high-correlated input variables were used to predict the TEY after the elimination of the low-correlated. The eight input variables remained for the CO prediction after the removal of the TAT and AP, while the NO_x prediction retained the nine input variables subsequent to the deletion of the TAT. The results of the model evaluation indicate that the TEY accuracy was the most accurate with MSE, MAE and MAPE values of 0.008, 0.71, and 0.006, respectively. The NO_x accuracy with an MSE of 6.148 indicated that the predicted value of NO_x differed from the actual value by ±6.148. With the MAPE value of 0.096, the NO_x accuracy revealed that a discrepancy between the expected and actual values of NO_x of ±9.6 % was observed. According to the NO_x accuracy with an MSE of 69.479, the predicted value of NO_x differs from the actual value by 8.335.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This work was financially supported by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok (Grant No. KMUTNB-FF-66-60).

Data availability

Manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Acknowledgements

The authors appreciate the support and infrastructure provided by the Division of Instrumentation and Automation Engineering Technology, Faculty of Engineering and Technology, King Mongkut's University of Technology North Bangkok (Rayong Campus) for the completion of this work. The authors gratefully acknowledge the financial support provided by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok (Grant No. KMUTNB-FF-66-60).

References

[1] Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z. et al. (2010). Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. Atmospheric Chemistry and Physics, 10 (15), 7017–7039. https://doi.org/10.5194/acp-10-7017-2010

[2] Amann, M., Kiesewetter, G., Schöpp, W., Klimont, Z., Winiwarter, W., Cofala, J. et al. (2020). Reducing global air pollution: the scope for further policy interventions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 378 (2183), 20190331. https://doi.org/10.1098/rsta.2019.0331

[3] Balali-Mood, M., Ghorani-Azam, A., Riahi-Zanjani, B. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. Journal of Research in Medical Sciences, 21 (1), 65. https://doi.org/10.4103/1735-1995.189646

[4] Larki, I., Zahedi, A., Asadi, M., Forootan, M. M., Farajollahi, M., Ahmadi, R., Ahmadi, A. (2023). Mitigation approaches and techniques for combustion power plants flue gas emissions: A comprehensive review. Science of The Total Environment, 903, 166108. https://doi.org/10.1016/j.scitotenv.2023.166108

[5] Okafor, E. C., Somarathne, K. D. K. A., Ratthanan, R., Hayakawa, A., Kudo, T., Kurata, O. et al. (2020). Control of NO_x and other emissions in micro gas turbine combustors fuelled with mixtures of methane and ammonia. Combustion and Flame, 211, 406–416. https://doi.org/10.1016/j.combustflame.2019.10.012

[6] Andrews, G. E. (2013). Ultra-low nitrogen oxides (NO_x) emissions combustion in gas turbine systems. Modern Gas Turbine Systems, 715–790. https://doi.org/10.1533/9780857096067.3.715

[7] Jain, S., Kaur, N., Verma, S., Kavita, Hosen, A. S. M. S., Sehgal, S. S. (2022). Use of Machine Learning in Air Pollution Research: A Bibliographic Perspective. Electronics, 11 (21), 3621. https://doi.org/10.3390/electronics11213621

[8] Kumar, V., Singh, S., Ahmed, Z., Verma, N. (2022). Air pollution prediction using machine learning algorithms: A systematic review. International Journal of Engineering Research & Technology, 11 (12), 84–93. Available at: https://www.ijert.org/research/air-pollution-prediction-using-machine-learning-algorithms-a-systematic-review-IJERTV11IS120026.pdf

[9] Deepu, B. P., Rajput, R. P. (2022). Air pollution prediction using machine learning. International Research Journal of Engineering and Technology, 9 (7), 2267–2271. Available at: https://www.irjet.net/archives/V9/i7/IRJET-V9I7416.pdf

[10] Ji, Y., Zhi, X., Wu, Y., Zhang, Y., Yang, Y., Peng, T., Ji, L. (2023). Regression analysis of air pollution and pediatric respiratory diseases based on interpretable machine learning. Frontiers in Earth Science, 11. https://doi.org/10.3389/feart.2023.1105140

[11] Gas Turbine CO and NO_x Emission Data Set. Available at: https://archive.ics.uci.edu/dataset/551/gas+turbine+co+and+nox+emission+data+set

[12] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24 (3), 69–71. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/pdf/MMJ2403-0069.pdf

[13] Lee, K., Im, S., Lee, B. (2023). Prediction of renewable energy hosting capacity using multiple linear regression in KEPCO system. Energy Reports, 9, 343–347. https://doi.org/10.1016/j.egyr.2023.09.121

[14] Sahin, G., Isik, G., van Sark, W. G. J. H. M. (2023). Predictive modeling of PV solar power plant efficiency considering weather conditions: A comparative analysis of artificial neural networks and multiple linear regression. Energy Reports, 10, 2837–2849. https://doi.org/10.1016/j.egyr.2023.09.097

[15] Rekabi, S., Sazvar, Z., Goodarzian, F. (2023). A machine learning model with linear and quadratic regression for designing pharmaceutical supply chains with soft time windows and perishable products. Decision Analytics Journal, 9, 100325. https://doi.org/10.1016/j.dajour.2023.100325

Received date 14.11.2023
Accepted date 11.01.2024
Published date 31.01.2024

***How to cite:** Photsathian, T., Suttikul, T., Tangsrirat, W. (2024). Prediction of air pollution from power generation using machine learning. EUREKA: Physics and Engineering, 1, 27–35. doi: https://doi.org/10.21303/2461-4262.2024.003148*