SIMTITS
Interdisciplinary School of Management and Technology
Institut Teknologi Sepuluh Nopember

JURNAL TEKNOBISNIS
2023 Vol. 9(1), 38-48

Received 17 Mar, 2023; Revised 28 Apr, 2023; Accepted 19 May, 2023 | DOI: 10.12962/j24609463.v9i1.929

Modeling Prediction of Sulfur Dioxide (SO_2), Nitrogen Oxides (NO_X) Emissions and Particulate from Coal-Fired Power Plant Using Machine Learning

Rina Yani L. Gaol[1,2*], Mohamad Atok[3]

ABSTRACT

Coal-fired power plants remain a crucial source of electricity in meeting national energy demands, despite the environmental challenges posed by the emissions generated during coal combustion. Emission coal-fired power plants include Sulfur Dioxide, Nitrogen Oxides, particulate matter have negative impact to the environment. To mitigate these emissions, regular monitoring, and measurement of gas emissions, as well as the development of emission prediction models, are essential. Machine learning has emerged as a promising approach for predicting gas emissions. In this study, Gradient Boosting (GB), Artificial Neural Network (ANN), and Support Vector Regression (SVR) machine learning models are employed. Factors influencing SO2, NOx, and particulate matter emissions are obtained from Distributed Control System records of the boiler system, along with the gas emission control technologies, comprising a total of 19 independent variables. Optimization of the boiler system can contribute to the reduction of gas emissions. The analysis results, measured by RMSE, R Square, and MAE, indicate that Gradient Boosting yields the most accurate predictions for gas emissions and particulate matter.

KEYWORDS: Emission Prediction, Gradient Boosting, ANN, SVR

[1]Environment Superintendent, PT Kalimantan Aluminium Industry (Adaro Minerals), Gunung Mas, Indonesia
[2]Business Analytics, Interdisciplinary School of Management and Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Department of Technology Management, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*Corresponding author: rinsaja19@gmail.com

1. INTRODUCTION

Indonesia currently relies on coal fired power plants to meet its national electricity needs while reducing reliance on oil-based power generation. Coal is abundant and its use can alleviate the burden of electricity subsidies, which have been straining the national budget due to rising oil prices. Emissions resulting from the combustion of coal in power plants include Sulfur Dioxide, Nitrogen Dioxide, particulate matter/dust commonly labelled as air pollutants due to their detrimental effects on the environment. These emissions contribute to environmental impacts such as global warming and climate change, primarily due to excessive concentrations in the atmosphere (Irsyad et al., 2020) Predictive models for gas emissions can assist in determining emission reduction strategies, identifying influencing factors, and estimating future emission levels (Bangert, 2021).

Machine learning is one of the predictive methods that can be utilized to predict gas emissions. According to (Ray, 2019), machine learning is a technique that enables machines or programs to perform specific tasks by utilizing data as input. By leveraging machine learning algorithms, models can learn from historical emission data and associated factors to make predictions and provide valuable insights for emission management and mitigation strategies.

In this study, machine learning techniques, including Gradient Boosting, Artificial Neural Network, and Support Vector Regression, are applied to predict gas emissions from coal-fired power plants. The aim is to develop accurate prediction models that can help in understanding and addressing the challenges associated with coal-based electricity generation, ultimately contributing to sustainable energy practices and environmental protection.

2. LITERATURE REVIEW

Emission Control and Particulate Technologies

Emission control in coal-fired power plants can be achieved through the selection of appropriate boiler technologies. PT. XYZ, an Independent Power Producer (IPP) with a capacity of 100 MWh, is committed to reducing emissions by utilizing Circulating Fluidized Bed (CFB) Boiler technology. According to (Yu et al., 2021) CFB is a clean coal combustion technique that has made significant advancements in the past five decades. (Krzywanski & Nowak, 2016) suggested Artificial Neural Network (ANN) models to predict SO2 emissions based on combustion conditions, air and oxygen mixture. The variables used in the ANN models included bed temperature, oxygen content in the air and oxygen mixture, humidity, fuel feed rate, and the type of fuel used. Additionally, limestone injection, a simple and cost-effective Flue Gas Desulfurization (FGD) technology, was utilized to remove SO2 from the flue gas (L. K. Wang et al., 2004).

(C. Wang et al., 2018) suggested employed Gaussian Process (GP) models to optimize coal combustion in boilers and reduce NOx emissions. The independent variables used in the GP models included boiler process variables, fuel composition, and

boiler geometry variables. Other factors influencing the reduction of NOx emissions were boiler load (MW), secondary air velocity (m/s), secondary air temperature (°C), oxygen concentration, and flue gas temperature (°C) (Yang et al., 2020). ESP is a particle control device used to remove fine particles from gas streams. (Peng et al., 2022) demonstrated that SVR techniques accurately predicted the concentration of 2.5-micron- sized particles.

Prediction Model

ANN (Artificial Neural Network) is widely used for developing prediction models that can establish accurate relationships between input and output in nonlinear systems. The ANN model consists of input neuron layers (nodes or units, which can range from one to several) hidden neuron layers, and output neuron layers (Zakaria et al., 2014) It utilizes interconnected nodes that mimic the structure and function of biological neurons to process and transmit information, enabling the model to learn complex patterns and relationships in the data. Boosting, as described by (Si & Du, 2020), is an ensemble method that combines multiple weak learners to generate a strong learning model. In boosting, weak learners are iteratively trained, and each subsequent learner focuses on correcting the mistakes made by the previous learners, ultimately leading to an accurate prediction model. Support Vector Regression (SVR) is a regression algorithm similar to Support Vector Machine (SVM) but designed for regression tasks (Smola & Scholkopf, 2004). SVR utilizes the concept of kernel methods, which allows for geometric interpretation of the learning algorithm in a nonlinear feature space.

3. METHODS

At this point, the method used in this study will be described, such as Data Collection, Preprocessing Data, Prediction Model and Evaluation

Data Collection

Data collection is secondary data, specifically historical records obtained from the Distributed Control System (DCS) sensors installed in the stacks of PT.XYZ, spanning a one-year period from June 2021 to June 2022. The study incorporates two types of variables: dependent variables and independent causal variables. The dependent variables in this research consist of three measurements: SO2, NOx, and particulate emissions recorded in Continuous Emission Monitoring Systems (CEMS), expressed in mg/Nm3. The selection of independent variables or factors influencing SO2, NOx, and particulate emissions was based on previous literature studies, which are presented in Table 1.

TABLE 1. Independent Variabel

No	Variable	Unit	Definition
1	CFR	ton/hr	Coal Flow Rate
2	MLD	Mw	Main Load
3	FUP	MPa	Furnace Upper Pressure
4	FUT	OC	Furnace Temperature
5	TSD	RPM	Turbine Speed

No	Variable	Unit	Definition
6	MSF	t/h	Main Steam Flow
7	MSP	MPa	Main Steam Pressure
8	MST	OC	Main Steam Temperature
9	RHP	MPa	Reheated Steam Pressure
10	RHT	MPa	Reheated Steam Temperature
11	CVP	MPa	Condition Vacuum Pressure
12	TAF	KNm3/h	Total Air Flow
13	BAT	OC	Bed Average Temperature
14	ECT	OC	Economizer Temperature
15	CYP	MPa	Cyclone Pressure
16	PFP	MPa	Primary Air Fan Pressure
17	SFP	MPa	Secondary Air Fan Pressure
18	LFR	-	Limestone Feed Frequency
19	O2C	%	O2 Content

Data Preprocessing for Missing Value

The missing value in this study can include outliers and errors resulting from sensor disruptions that prevent the reading and recording of data. To address the missing values in the obtained data, linear interpolation method was employed. Linear interpolation is a data processing technique that estimates missing values by following the same pattern as the surrounding available data. It can be applied to handle missing value in power generation facilities (Wang et al., 2023).

Data Normalization

Data normalization refers to the process of transforming data to a specific scale with the aim of enhancing the model's performance in predicting data, managing outliers, and reducing data redundancy (Singh & Singh, 2022). The normalization technique to be employed in this study is Min-Max (MM) normalization.

Feature Identification and Selection

During the data processing stage, correlation analysis and feature selection will be conducted. The correlation analysis utilized will be Pearson correlation, which aims to measure the linear relationship between two variables.

Prediction Analysis

Before conducting machine learning prediction, a classification of the dependent variable will be performed. The classification method employed is regression. Through this initial classification, considering factors from the available data is expected to enhance the accuracy of gas and particulate emission predictions. The machine learning algorithms utilized in this study are Gradient Boosting, ANN, and SVR. The prediction model for Gradient Boosting adopts a Gaussian distribution, as the variables are continuous, and the number of trees built is set to 1000. Parameters respectively such as shrinkage and interaction depth are adjusted to 0.01 and 4. In all three methods, data validation will be performed through a test: train data split ratio of 80:20.

Model Performance Evaluation

To ensure the effectiveness of the prediction model using the training data, a specific evaluation will be conducted. The model's performance on the training data will be assessed quantitatively using root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R2). RMSE and MAE serve as evaluation metrics to estimate how far the regression model's predicted values deviate from the observed values. RMSE and MAE are used to measure the prediction errors in the same units as the dependent variable. RMSE is more sensitive to outliers or values far from the true values, while MAE is more sensitive to small errors between predicted and true values (James et al., 2017)).

In evaluating the prediction model, the following conditions indicate the optimal model selection:

- If the RMSE value is smaller, the model's performance is better because it indicates that the predicted values are closer to the actual values.
- If the MAE value is smaller, the model's performance is better because it indicates that the difference between the predicted values and the actual values is smaller.

4. RESULTS

Datasets

The total number of collected numerical data in this study is 9480. The statistical description of the dependent variables is as follows:

TABLE 2. Dependent Variable Characteristic Data

Variable	Min	Max	Mean	Standard Deviation
SO_2	0	1,707.030	230.903	129.003
NOX	0	143.604	87.519	34.080
Partikulat	21.867	40.520	25.424	2.109

As shown in Table 3, the minimum values of 0 for SO2, NOX represent negative recorded values in the DCS. These values indicate that the boiler was not in operation during those instances, and thus the values were adjusted to 0. Additionally, the standard deviation values for all the variables are smaller than their respective means, indicating that the data distribution for all the dependent variables is centered around the mean value. Descriptive statistics for the independent variables are presented in the following table:

TABLE 3. Independent Variable Characteristic Data

Variable	Min	Max	Mean	Standard Deviation
CFR	0	100.528	54.363	20.461
MLD	0.019	116.284	67.449	26.984
FUP	-376.406	365.625	-73.961	57.549
FUT	26.881	824.441	626.995	185.536
TSD	0	3050.020	2719.850	881.234

Variable	Min	Max	Mean	Standard Deviation
MSF	0	368.592	211.703	83.262
MSP	0.001	13.833	8.715	3.200
MST	26.269	543.747	484.957	133.062
RHP	0.033	3.014	1.804	0.679
RHT	26.900	543.389	473.473	132.865
CVP	-93.102	0.609	-82.881	25.737
TAF	0	351.413	220.912	71.316
BAT	25.386	961.283	713.531	201.266
ECT	25.791	349.406	261.232	69.769
CYP	-0.064	1.611	0.611	0.324
PFP	-0.460	16.613	11.968	1.406
SFP	-0.570	13.348	3.870	3.652
LFR	0.007	10.310	0.366	2.788
CFR	0	100.528	54.363	20.461
O2C	0.132	24.080	4.583	4.520

Descriptive Statistical Results as shown in Table 3 indicate that a value of 0 represents the minimum value for variables CFR, TSD, MSF and TAF. These values indicate the reading conditions in the DCS (Distributed Control System) when the boiler unit is not in operation. On the other hand. The negative minimum values for the dependent variable pressure in FUP, CVP, CYP, PFP and SFP imply that the equipment operates under negative pressure, which is below atmospheric pressure. The standard deviation values for all independent variables, except LFR are smaller than their respective means. This suggests that the data distribution for all these dependent variables is centered around the mean value. However, the larger standard deviation values for LFR indicates significant variations in the data, with a wider spread from the mean.

Missing Value

In this study, a thorough examination of missing values was conducted for each variable. Additionally, categorical null values recorded in the data were converted or replaced as missing data. The results of the missing value analysis are presented in Table 4.

TABLE 4. Missing Value Results Analysis

No	Variable	Missing Value	%
1	SO2	4	0.042%
2	NOX	2	0.021%
3	Particulate	3	0.032%
4	CFR	1	0.011%
5	MLD	1	0.011%
6	FUP	1	0.011%
7	FUT	1	0.011%
8	TSP	1	0.011%
9	MSF	1	0.011%
10	MSP	1	0.011%
11	MST	1	0.011%

No	Variable	Missing Value	%
12	RHP	1	0.011%
13	RHT	1	0.011%
14	CVP	3	0.032%
15	TAF	1	0.011%
16	BAT	1	0.011%
17	ECT	1	0.011%
18	CYP	1	0.011%
19	LFR	128	1.350%
20	PFP	1	0.011%
21	SFP	1	0.011%
22	O2C	1	0.011%

As shown in Table 4, it can be observed that the missing values for the three dependent variables, as well as the independent variables except for LFR have a very low percentage below 1%. These missing data points are considered to have minimal impact on the prediction performance (Kang, 2013). For LFR variable, missing values will be managed using interpolation techniques. and the same treatment will be applied to all variables with missing values.

Data Normalization

In normalizing the data in this study, as an example with a dust value of 25.781. The minimum and maximum values for the dust variable are found to be 21.867 and 40.52. Therefore, the normalized value is calculated as follows:x_norm=(x_obs-x_min) / (x_max-x_min) = (25.781- 21.867) / (40.52-21.867) = 0.209

Feature Identification and Selection

The selection of influential variables is performed by calculating the correlation values between each independent and dependent variable. as well as the p-value < 0.05. Table 5 presents the p-value results for the correlation between dependent and independent variables indicating significant values. The selection of independent variables is determined based on correlation values less than 0.3 or greater than -0.3. Variables that do not meet these criteria are considered weak, and they are not included in the model. The independent variable SO2 does not include SFP features in the model. For NOX, all independent variables are significant. Regarding Particulate variable, the features not included in the model are FUP, TAF, CYP, LFR and SFP.

TABLE 5. Correlation and P value Results

No	Variable	SO_2		NOX		Particulate	
		Correlation	Pvalue	Correlation	Pvalue	Correlation	Pvalue
1	CFR	0.541	0	0.791	0	-0.371	0
2	MLD	0.489	0	0.764	0	-0.337	0
3	FUP	-0.321	0	-0.512	0	0.135	0
4	FUT	0.576	0	0.884	0	-0.413	0
5	TSP	0.579	0	0.870	0	-0.457	0
6	MSF	0.490	0	0.772	0	-0.361	0

No	Variable	SO_2		NOX		Particulate	
		Correlation	Pvalue	Correlation	Pvalue	Correlation	Pvalue
7	MSP	0.513	0	0.808	0	-0.325	0
8	MST	0.575	0	0.866	0	-0.376	0
9	RHP	0.494	0	0.788	0	-0.373	0
10	RHT	0.567	0	0.857	0	-0.354	0
11	CVP	-0.577	0	-0.870	0	0.369	0
12	TAF	0.507	0	0.787	0	-0.260	0
13	BAT	0.582	0	0.885	0	-0.403	0
14	ECT	0.558	0	0.855	0	-0.349	0
15	CYP	0.400	0	0.621	0	-0.241	0
16	LFR	0.420	0	0.082	1.33E-15	-0.075	3.59E-13
17	PFP	0.566	0	0.836	0	-0.368	0
18	SFP	0.276	0	0.462	0	-0.049	1.65E-06
19	O2C	-0.572	0	-0.875	0	0.452	0

Data Analysis

Before making predictions on the data. regression analysis will be conducted to determine which variables influence each dependent variable. As shown in Table 6, the R-square value for SO2 is found to be 0.58, indicating that approximately 58% of the SO2 variable can be explained by the independent variables. The R-square value obtained for the NOx variable is 0.8672, meaning that around 86.72% of the NOX variable can be explained by the independent variables. The R- square value for particulate matter is 0.6611, indicating that approximately 66.11% of the particulate variable can be explained by the independent variables. Regarding the coefficient values, if a variable has a positive coefficient, it means that the variable has a positive influence on SO2. Conversely, if a variable has a negative coefficient, it indicates a negative influence on the dependent variable.

TABLE 6. Regression Results

Dependent Variable	Independent Variable	Coeficient	P-Value	R Square
SO_2	CFR	0.4746	< 2e-16	0.58
	MLD	0.7041	< 2e-16	
	FUP	-0.0215	2.04E-02	
	TSD	0.1142	< 2e-16	
	MSF	-0.8597	< 2e-16	
	MSP	-0.0796	3.15E-05	
	RHP	-0.1122	1.86E-04	
	TAF	-0.2247	< 2e-16	
	BAT	0.0656	2.88E-08	
	CYP	0.0721	1.01E-05	
	LFR	0.1760	< 2e-16	
	PFP	0.1395	< 2e-16	
	O2C	0.1149	1.06E-08	
NO_X	CFR	-0.517	< 2e-16	0.8672
	MLD	0.771	< 2e-16	

Dependent Variable	Independent Variable	Coeficient	P-Value	R Square
	FUT	0,583	< 2e-16	
	TSD	0.205	< 2e-16	
	MSF	-1.047	< 2e-16	
	MSP	0.150	< 2e-16	
	MST	-0.245	1,72E-08	
	RHT	-0.092	3.61E-08	
	CVP	-0.141	0.0144	
	TAF	0.894	2.49E-15	
	BAT	0.217	< 2e-16	
	CYP	0.240	< 2e-16	
	PFP	-0.607	< 2e-16	
	SFP	-0.320	< 2e-16	
	O2C	-0.276	< 2e-16	
Particulate	CFR	0.305	<2e-16	0.6611
	MLD	1.648	<2e-16	
	FUT	-0.910	<2e-16	
	TSP	-0.690	<2e-16	
	MSF	-1.619	<2e-16	
	MSP	0.209	<2e-16	
	MST	-0.949	<2e-16	
	RHP	-0.577	<2e-16	
	RHT	1.204	<2e-16	
	CVP	-0.281	<2e-16	
	BAT	0.223	<2e-16	
	ECT	0.520	<2e-16	
	PFP	0.307	<2e-16	
	O2C	0.349	<2e-16	

Prediction Model Analysis

In evaluating the best predictive model among Gradient Boosting, Artificial Neural Network (ANN), and Support Vector Regression (SVR), a comparison is from analysis results of Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

TABLE 7. Regression Results

Variable	Method	MSE	RMSE	MAE
SO_2	Gradient	0.002	0.05	0.039
	ANN	0.004	0.066	0.051
	SVR	0.004	0.063	0.048
NO_X	Gradient	0.004	0.066	0.051
	ANN	0.01	0.1	0.084
	SVR	0.007	0.087	0.071
Particulate	Gradient	0.012	0.109	0.1
	ANN	0.018	0.133	0.117
	SVR	0.013	0.112	0.1

As shown in Table 7, it can be observed that for SO2, NOx and Particulate, Gradient Boosting is selected as the best performance predicting model, as indicated by lower RMSE and MAE values compared to other methods.

5. CONCLUSIONS

Based on the correlation analysis, the factors influencing the formation of SO2, NOx and particulate gas emissions are as follows:

- The factors affecting SO2 gas emissions are Coal Flow Rate, Main Load, Furnace Upper Pressure, Turbine Speed, Main Steam Flow, Main Steam Pressure, Reheated Steam Pressure, Total Air Flow, Bed Average Temperature, Cyclone Pressure, Limestone Feed Frequency, Primary Air Fan Pressure, O2 Content with an R-squared value of 58%. This indicates that 58% of the variation in SO2 emissions can be explained by these variables.
- The factors influencing NOx gas emissions are Coal Flow Rate, Main Load, Furnace Upper Temperature, Turbine Speed, Main Steam Flow, Main Steam Pressure, Main Steam Temperature, Reheated Steam Pressure, Condition Vacuum Pressure, Total Air Flow, Bed Average Temperature, Cyclone Pressure, Primary Air Fan Pressure, Secondary Air Fan Pressure, O2 Content with an R-squared value of 86,72%. This implies that 86.72% of the variation in NOx emissions can be explained by these variables.
- The factors influencing particulate matter are Coal Flow Rate, Main Load, Furnace Upper Temperature, Turbine Speed, Main Steam Flow, Main Steam Pressure, Main Steam Temperature, Reheated Steam Pressure, Reheated Steam Temperature, Condition Vacuum Pressure, Bed Average Temperature, Economizer Temperature, Primary Air Fan Pressure, O2 Content with an R-squared value of 66.11%. This indicates that 66.11% of the variation in particulate matter can be explained by these factors/variables.

The results of machine learning techniques evaluation are as follows:

- The best performing model in predicting SO2, based on the analysis of RMSE and MAE is Gradient Boosting.
- The best performing model in predicting NOx, based one analysis of RMSE and MAE is Gradient Boosting.
- The best performing model in predicting particulate matter based on analysis of RMSE and MAE is Gradient Boosting.

REFERENCES

Bangert, P. (2021). Machine Learning and Data Science in the Oil and Gas Industry Best Practices, Tools, and Case Studies. In *Machine Learning and Data Science in the Oil and Gas Industry*. Elsevier. https://doi.org/10.1016/b978-0-12-820714-7.00020-0

Irsyad, M. I. Al, Halog, A., Nepal, R., & Koesrindartoto, D. P. (2020). Economical and environmental impacts of decarbonisation of Indonesian power sector. *Journal of*

Environmental Management, *259*, 1–11. https://doi.org/10.1016/j.jenvman.2019.109669

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning : with applications in R* (Vol. 8). Springer New York Heidelberg Dordrecht London.

Kang, P. (2013). Locally linear reconstruction based missing value imputation for supervised learning. *Neurocomputing*, *118*, 65–78. https://doi.org/10.1016/j.neucom.2013.02.016

Krzywanski, J., & Nowak, W. (2016). Artificial Intelligence Treatment of SO2 Emissions from CFBC in Air and Oxygen-Enriched Conditions. *Journal of Energy Engineering*, *142*(1), 1–10. https://doi.org/10.1061/(asce)ey.1943-7897.0000280

Peng, J., Han, H., Yi, Y., Huang, H., & Xie, L. (2022). Machine learning and deep learning modeling and simulation for predicting PM2.5 concentrations. *Chemosphere*, *308*, 1–8. https://doi.org/10.1016/j.chemosphere.2022.136353

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)*, 35–39.

Si, M., & Du, K. (2020). Development of a predictive emissions model using a gradient boosting machine learning method. *Environmental Technology and Innovation*, *20*, 1–17. https://doi.org/10.1016/j.eti.2020.101028

Smola, A. J., & Scholkopf, B. (2004). A tutorial on support vector regression *. *Statistics and Computing*, *14*, 199–222.

Wang, C., Liu, Y., Zheng, S., & Jiang, A. (2018). Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process. *Energy*, *153*, 149–158. https://doi.org/10.1016/j.energy.2018.01.003

Wang, L. K., Pereira, N. C., & Hung, Y.-T. (2004). *Air Pollution Control Engineering*. Humana Press Inc.

Yu, H., Gao, M., Zhang, H., & Chen, Y. (2021). Dynamic modeling for SO2-NOx emission concentration of circulating fluidized bed units based on quantum genetic algorithm - Extreme learning machine. *Journal of Cleaner Production*, *324*, 1–14. https://doi.org/10.1016/j.jclepro.2021.129170

Zakaria, M., Al-Shebany, M., & Sarhan, S. (2014). Artificial Neural Network : A Brief Overview. *Journal of Engineering Research and Applications*, *4*(2), 7–12. www.ijera.com

How to cite this article:

Gaol, R. Y. L., & Atok, M. (2023). Modeling Prediction Of Sulfur Dioxide (So2), Nitrogen Oxides (Nox) Emissions And Particulate From Coal- Fired Power Plant Using Machine Learning. *Jurnal Teknobisnis, 9*(1): 38-48. DOI: 10.12962/j24609463.v9i1.929