IJECBE

IJECBE (2024), **2**, **4**, 454–476
Received (4 July 2024) / Revised (1 July 2024)
Accepted (5 July 2024) / Published (30 December 2024)
https://doi.org/10.62146/ijecbe.v2i4.77
https://ijecbe.ui.ac.id
ISSN 3026-5258

International Journal of Electrical, Computer and Biomedical Engineering

RESEARCH ARTICLE

Optimization of Heat Rate and Greenhouse Gas Emission Reduction at Coal-Fired Power Plants in Indonesia Through Machine Learning Modeling

Ariandiky Eko Setyawan* and Budi Sudiarto

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia
*Corresponding author. Email: ariandiky.setyawan@pln.co.id

Abstract

This study aims to develop predictive models for the heat rate of coal-fired steam power plants (CFSPPs) in Indonesia using various machine learning techniques and to identify factors influencing greenhouse gas emissions, specifically CO_2. Techniques used include Linear Regression, Lasso Regression, Polynomial Regression, Ridge Regression, Support Vector Regression, Random Forest Regression, Gradient Boosting Regression, Elastic Net Regression, AdaBoost Regression, Neural Network Regression, Decision Tree Regression, and Extra Trees Regression. The data consists of 468 performance test results from CFSPPs, covering operational parameters such as boiler type, ambient temperature, flue gas temperature, and unburned carbon. The correct values should be: "Analysis shows that the Extra Trees Regression model provides the best performance with an R-squared value of 0.946, MAE of 136.884, MSE of 35053.258, and RMSE of 187.225 for heat rate modeling, and an R-squared value of 0.992, MAE of 20.143, MSE of 1494.351, and RMSE of 38.657 for CO_2 emissions modeling, demonstrating high accuracy and good generalization. Significant factors influencing the heat rate include Gross Power Output (GPO), Net Power Output (NPO), load percentage, boiler type, coal HHV, coal consumption, and operational duration. This model is implemented using the Postman application for realtime heat rate and CO_2 emissions prediction, facilitating integration with CFSPP's operational systems. The research results indicate that the application of machine learning can improve energy efficiency and reduce CO_2 emissions, supporting Indonesia's Nationally Determined Contribution (NDC) targets. This study provides new insights into the application of machine learning in the power generation industry and offers recommendations for further implementation and research.

Keywords: machine learning, heat rate, greenhouse gas emissions, extra trees regression

1. Introduction

Coal-fired steam power plants (CFSPPs) are essential in meeting Indonesia's ever-increasing energy demands. Powerplants contribute significantly to the national power generation capacity, with more than 50% of electricity generated from these fossil-fuel-based plants [1] [2]. However, the heavy reliance on fossil fuels leads to significant greenhouse gas emissions, particularly carbon dioxide (CO_2). These emissions have detrimental environmental impacts, contributing to global climate change and air pollution [3] [4]. Therefore, there is an urgent need to enhance the operational efficiency of powerplant's to mitigate their environmental impact [5].

Heat rate is a key parameter in measuring the efficiency of power plants. It indicates the amount of fuel energy required to produce one unit of electricity. A lower heat rate signifies higher efficiency, meaning less fuel is needed to generate the same amount of electricity [6] [7]. Thus, optimizing the heat rate can result in significant fuel savings and reduced CO_2 emissions. Efforts to improve this efficiency are crucial to minimizing the environmental impact of CFSPP's [7].

Machine learning technology offers innovative solutions to tackle these challenges [2]. With the ability to analyze vast amounts of operational data, machine learning can identify patterns and trends that traditional methods might overlook [8]. Machine learning models can be used to predict heat rate and CO_2 emissions with high accuracy, enabling better decision-making and effective optimization strategies [9]. By leveraging historical and real-time operational data, these models can provide actionable recommendations to enhance operational efficiency and reduce greenhouse gas emissions [10].

Indonesia has committed to reducing greenhouse gas emissions in accordance with the Paris Agreement. However, data from the Ministry of Environment and Forestry indicates that emissions from the energy sector, particularly CFSPPs, remain a major contributor [1] [11]. To achieve the Nationally Determined Contribution (NDC) targets, concrete steps are needed to improve the operational efficiency of CFSPP's [1] [3]. This research aims to develop predictive models using machine learning techniques that can identify key factors affecting heat rate and CO_2 emissions and propose effective solutions for improvement [10] [11].

The study focuses on developing predictive models for heat rate and CO_2 emissions using performance test data from CFSPPs conducted by companies that perform unit performance testing [10]. The data includes various testing parameters such as Gross Power Output (GPO) and Net Power Output (NPO), load percentage, boiler type, coal HHV, coal consumption, and operational duration [12] [13]. Consequently, this research is expected to make a significant contribution to enhancing energy efficiency, as indicated by the heat rate values, and reducing the environmental impact of CFSPP's operations, particularly for CO_2 emission parameters [6] [2]. The findings of this study are also anticipated to provide a basis for the development of further policies and strategies in sustainable and environmentally friendly energy management [1] [14].

2. Plant Operations and Machine Learning Modeling)

2.1 CFSPPs Operational Process

CFSPPs begins with heating water in the boiler to generate main steam. Initially, feedwater is fed into the condenser and pumped by the condensate pump through the LP heater to the deaerator to remove oxygen. Next, this water is pumped by the boiler feed pump into the economizer. The water flows through pipes and is heated in the boiler tubes. Fossil fuels such as coal, oil, or natural gas are burned in the boiler to produce heat. This combustion process involves a chemical reaction between the fuel and oxygen from the air, producing heat, exhaust gases (including CO_2, $NO_\times$, and SO_2), and ash [15]. Efficient combustion is crucial to maximize the energy obtained from the fuel and minimize air pollution [16].

The heat generated from fuel combustion is used to heat the water in the boiler, producing high-pressure, high-temperature steam. This steam is then collected in the steam drum and further heated in the superheater to become high-pressure, dry steam. The high-pressure steam enters the turbine and rotates the turbine blades [7]. The quality of the produced steam, including its pressure and temperature, is critical for turbine efficiency. The thermal energy from the steam is converted into mechanical energy used to spin the turbine rotor. The turbine consists of several stages, including the high-pressure (HP) turbine, intermediate-pressure (IP) turbine, and low-pressure (LP) turbine. Each stage is designed to optimize the utilization of steam energy at various pressure levels.

The rotation of the turbine shaft drives the generator shaft, which is connected via a coupling. The electrical generator spun by the turbine produces electrical energy. The generator operates on the principle of electromagnetic induction, where the magnetic field generated by the rotating rotor cuts through the wire windings on the stator, generating an electric current. This electrical energy is then distributed through the distribution network for use by consumers. After passing through the turbine, the steam, which has lost most of its energy, is condensed back into water in the condenser. This condensate water is then pumped back along with additional make-up water to the boiler for the next cycle. An efficient condensation system is essential to reduce heat loss and ensure the continuous circulation of water within the system [13] [17].

2.2 Powerplant performance indicator

2.2.1 Gross Power Output (GPO) and Net Power Output (NPO)

Gross Power Output is the total amount of power produced by the power plant, while Net Power Output is the power available for distribution after accounting for the plant's internal consumption. High values of GPO and NPO are indicative of a well-performing power plant [6].

2.2.2 Heat Rate:

Heat rate is a measure of the thermal efficiency of a power plant, indicating the amount of input energy required to produce one unit of output energy. The lower the heat rate value, the higher the efficiency of the power plant [18]. The formula for heat rate is:

$$Heatrate = \frac{Q_{in}}{Q_{out}} \tag{1}$$

2.2.3 Boiler efficiency

Boiler efficiency measures how effectively the boiler converts fuel energy into steam energy. High boiler efficiency indicates that the boiler is operating well with minimal energy lost as waste heat [16]. Boiler efficiency can be expressed as:

$$\eta_{boiler} = \frac{Q_{boiler\ output}}{Q_{fuel}} \tag{2}$$

2.2.4 Turbine efficiency

Turbine efficiency measures how effectively the turbine converts steam energy into mechanical energy. An efficient turbine generates more electrical energy from the same amount of steam [18]. Turbine efficiency can be expressed as:

$$\eta_{turbine} = \frac{Q_{mechanical}}{Q_{steam}} x100\% \tag{3}$$

2.2.5 Availability Factor:

The availability factor is the percentage of time the plant is available for full operation. A plant with a high availability factor has longer operational hours and less downtime [6]. The availability factor can be expressed as:

$$AF = \frac{T_{operation}}{T} x100\% \tag{4}$$

2.2.6 CO_2 Emissions per Unit Energy Output

This indicator measures the environmental impact of CFSPPs operations. CO_2 emissions are the primary greenhouse gas produced from the combustion of fossil fuels, and this emission rate is a critical indicator for assessing the environmental impact of the power plant. These performance indicators are essential for understanding and optimizing the operation of coal-fired power plants. By closely monitoring and improving these indicators, power plant operators can enhance efficiency, reduce fuel consumption, and minimize greenhouse gas emissions, contributing to more sustainable and environmentally friendly energy production.

2.3 Relationship Between Energy Efficiency and Greenhouse Gas Emissions

Energy efficiency and greenhouse gas emissions have a close relationship [15]. Higher energy efficiency leads to lower fuel consumption, which means reduced greenhouse gas emissions. For instance, improving combustion efficiency in the boiler can decrease the amount of fuel required to produce the same amount of energy, thus reducing CO_2 emissions. The CO_2 emissions are calculated using the formula [4]:

$$E_{CO_2} = DAxFE \tag{5}$$

where: E_{CO_2} : Total CO_2 emissions (tons CO_2) DA : Activity Data (TJ) FE : Emission Factor (tons/TJ) The activity data represents the consumption data per type of fuel that has been converted to energy units. The formula for converting fuel consumption data from mass units (tons) to energy units (TJ) is as follows:

$$DA_{BB} = F_{BB}xNCVx10^{-3} \tag{6}$$

where: DA_{BB} : Coal Activity Data (TJ) F_{BB} : Annual coal consumption (tons) NCV : Net calorific value of coal (TJ/Gg), the default national NCV values for coal

2.4 Machine Learning Techniques in Heat Rate Modeling

Machine learning is a branch of artificial intelligence that enables systems to learn from data and make predictions or decisions without being explicitly programmed [19] [20] [21]. In the context of heat rate prediction for coal-fired steam power plants (CFSPP's), machine learning techniques can be used to analyze operational data and identify complex patterns, which can help optimize energy efficiency and reduce greenhouse gas emissions.

Here are some common machine learning techniques used in heat rate prediction [9] [21]:

1. Linear regression is a statistical technique used to model the relationship between a dependent variable (y) and one or more independent variables (x). In the context of heat rate prediction, linear regression can be used to predict the heat rate based on variables such as plant load, steam pressure, and temperature. The formula for linear regression is:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{7}$$

   Where y is dependent variable (heat rate), x is independent variable (e.g., fuel consumption), β_0 is intercept, β_1 is regression coefficient and ε is error term.
2. Polynomial regression is a statistical method used to model the relationship between the dependent variable and the independent variables as a polynomial of a certain degree. The formula for polynomial regression is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x_n + \varepsilon \tag{8}$$

   where $x_1, x_2, \cdots, x_n$ are the independent variables.

3. Ridge regression is a regression technique used to analyze data that has multi-collinearity (high correlation between independent variables). The formula for ridge regression is:

$$minimize = \left(\sum_{i=1}^{n} \left(\gamma_1 - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \quad (9)$$

where β is the regularization parameter that determines the extent of the penalty applied to the regression coefficients.

4. Random Forest Regression is an ensemble learning technique used to improve prediction accuracy and reduce overfitting by combining multiple decision trees. It constructs a large number of decision trees during training and outputs the mean prediction (regression) of the individual trees. The formula for Random Forest Regression is:

$$\hat{y}(x) = \frac{1}{N} \sum_{i=1}^{N} h_i(x) \quad (10)$$

where $\hat{y}$ is the final prediction, and $h_i(x)$ is the prediction from the i -th tree for sample x.

5. Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression tasks. The formula for SVM regression is:

$$min_{w,b,\xi} = \left(\frac{1}{2} |0w| \, 0^2 + C \sum_{i=1}^{n} \xi_1 \right) \quad (11)$$

where w is the vector of coefficients, b is the intercept, ξ is the error variable, and C is the regularization parameter.

6. Gradient Boosting Regression is part of the boosting family of algorithms. Boosting is an ensemble technique that aims to improve model accuracy by combining several weak learners into a strong learner. The formula for gradient boosting regression is:

$$\hat{y} = \sum_{m=1}^{M} \gamma_m h_m(x) \quad (12)$$

where $\hat{y}$ is the final prediction, M is the total number of base models, $h_m(x)$ is the mmmth base model, and γ_m is the weight or coefficient of the m -th model.

7. Elastic Net Regression is a linear regression technique that combines L1 (Lasso) and L2 (Ridge) regularization to address some of the limitations of both methods. The formula for elastic net regression is:

$$min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (\gamma_i - X_i \beta)^2 + \lambda_1 \, |0\beta| \, 0_1 + \lambda_2 \, |0\beta| \, 0_2^2 \right) \quad (13)$$

where y_i is the target value, X_i is the feature vector, β is the vector of coefficients, $|0\beta|0_1$ is the L1 norm of the coefficients (Lasso penalty), $|0\beta|0_2^2$ is the L2 norm of the coefficients (Ridge penalty), and λ_1 and λ_2 are the regularization parameters controlling the strength of the L1 and L2 penalties.

8. AdaBoost Regression is an ensemble method used to improve the accuracy of base regression models. The final model is a combination of all the base models:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{14}$$

where $\hat{y}$ is the final prediction, M is the total number of base models, $_m(x)$ is the mmmth base model, and α_m is the weight or coefficient of the m -th model, indicating its contribution to the final prediction.

9. Neural Network Regression is an approach based on artificial neural networks used for regression tasks, i.e., predicting continuous values. The formula for neural network regression is:

$$y = f * (W_n f(W_{n-1}(\cdots f(W_1 * z + b_1)\cdots) + b_{n-1}) + b_n) \tag{15}$$

10. Extra Trees Regression is an ensemble learning method similar to Random Forest but with some key differences. Like Random Forest, Extra Trees build a large number of decision trees and combine their predictions to make a final prediction. However, Extra Trees use more random techniques to build decision trees, including the random selection of split points at each node.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \widehat{y_1} \tag{16}$$

2.5 Model Evaluation

Model evaluation is performed to assess the performance of the developed predictive models. The evaluation criteria include [22]:

1. Mean Absolute Error (MAE): MAE measures the average absolute error between predicted values and actual values. Lower MAE indicates a better model.

$$MAE = \frac{1}{N}\Sigma_{i=1}^{n} \left|y_i - \hat{y}_i\right| \tag{17}$$

2. Mean Squared Error (MSE): MSE measures the average squared error between predicted values and actual values. Lower MSE indicates a better model.

$$MSE = \frac{1}{N}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

3. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, providing a clearer interpretation of MSE in the same units as the original data.

$$RMSE = \sqrt{MSE} \tag{19}$$

4. R-squared (R^2): R^2 is a metric that measures the proportion of variability in the dependent variable that can be explained by the independent variables. R^2 values range from 0 to 1, with higher values indicating a better model.

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{20}$$

5. Mean Absolute Percentage Error (MAPE): MAPE is the average absolute percentage error, providing a perspective on the scale of the error relative to actual values.

$$MAPE = \frac{1}{n}\Sigma_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{21}$$

6. Median Absolute Percentage Error (MdAPE): MdAPE is the median of the absolute percentage errors, which is more robust to outliers compared to MAPE.

$$MdAPE = median\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\right) \tag{22}$$

2.6 Model Validation (K-Fold Cross-Validation)

Model validation is a crucial step in ensuring that the developed heat rate and CO_2 emissions prediction models perform well on unseen data. K-Fold Cross-Validation is a validation technique that divides the data into K subsets or "folds." The model is trained on K-1 folds and validated on the remaining fold. This process is repeated K times, with each fold serving as the validation set once. This technique helps to maximize the use of data and provides a more accurate estimate of the model's performance.

2.7 Model Implementation

The implementation of the heat rate prediction model is carried out by developing an API (Application Programming Interface) that allows users to send operational data from the CFSPP's and receive heat rate predictions. In this case, the simple implementation of the model can be done using the Postman application. This approach facilitates real-time integration of the model into the operational system of the power plant, enabling continuous monitoring and optimization of energy efficiency. The API can handle various data inputs, process them using the trained machine learning model, and return accurate predictions that help in making informed decisions to enhance the plant's performance and reduce greenhouse gas emissions.

3. Research Methods

As illustrated in the diagram below, this research procedure encompasses several stages, starting with a literature review. In this stage, a literature review is conducted to understand the basic concepts and relevant methods, including searching for journals, books, and other publications related to heat rate and CO_2 emissions prediction in power plant using machine learning techniques [13] [10][10]. Following this, problem identification is carried out to determine the main problem to be addressed in this study, which is the prediction of heat rate and CO_2 emissions in CFSPP's. The problem identification is done clearly to ensure the appropriate focus of the research [23] [2].

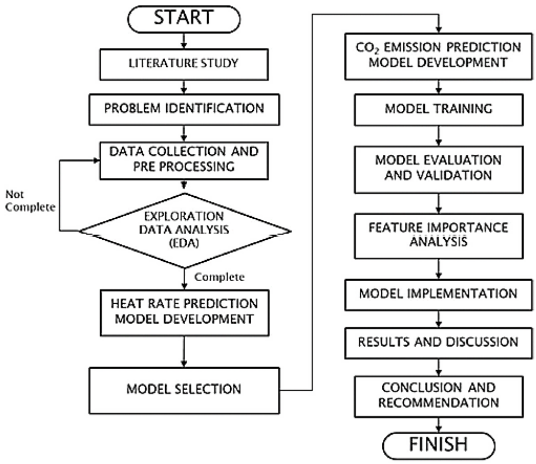

Figure 1. Research Flow Chart

Next, data collection is performed by gathering operational data from CFSPP's to be used in the model. This data includes various operational parameters such as load, temperature, humidity levels, and others that are relevant for predicting heat rate and CO_2 emissions. After the data is collected, data preprocessing is conducted to clean and transform the data for use in the model. This process includes handling missing data, encoding categorical variables, normalizing or standardizing features, and other preprocessing techniques to improve data quality.

Exploratory Data Analysis (EDA) is then conducted to understand the characteristics of the data and the relationships between variables. EDA helps identify patterns, anomalies, and important relationships between the features present in the data. If EDA is conducted thoroughly, it will provide deep insights into the data, which can help build better models. Conversely, if EDA is not done comprehensively, important information might be overlooked, negatively impacting the accuracy and performance of the developed models.

Following EDA, the development of heat rate prediction models is carried out by developing various machine learning models such as linear regression, polynomial regression, Lasso regression, Ridge regression, Support Vector Regression (SVR), Random Forest, Gradient Boosting Regression, Elastic Net Regression, AdaBoost Regression, and Neural Network Regression. The next stage is the selection of the best model based on evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The best-selected model is then used to develop the CO_2 emissions prediction model, considering relevant features.

Model training is conducted by splitting the data into training and testing sets to train the model and evaluate its performance. Model evaluation is performed using various evaluation metrics to measure how well the model predicts heat rate and CO_2 emissions. Model validation is also conducted to ensure its generalization capability, using validation techniques such as K-Fold Cross-Validation [17].

Feature importance analysis is performed to analyze the most influential factors in predicting heat rate and CO_2 emissions. Techniques such as SHAP values and feature importance plots are used to identify the features with the greatest impact on the model. Afterward, the model is implemented for real-time predictions using applications like Postman.

The final stage of this research is the results and discussion, where the obtained results are analyzed and discussed to understand their implications in the operational context of CFSPPs. Conclusions are drawn based on the analysis results, and recommendations are provided for further improvement or model application in real operational scenarios.

3.1 Data Collection

The data used in this study comes from performance tests conducted on various CFSPPs in Indonesia. These performance tests include comprehensive and crucial operational parameters essential for analyzing and optimizing power plant performance. These parameters include load, temperature, humidity levels, flue gas temperature, unburned carbon, coal moisture content, and various other relevant parameters.

These performance tests aim to provide an overall picture of the efficiency and performance of power plants under various operational conditions. The data collected from approximately 468 test records offers a very rich and diverse database, which is critical for developing reliable machine learning models. With this extensive and varied dataset, the research can more accurately predict heat rate and CO_2 emissions and identify key factors influencing CFSPPs performance [15].

Through in-depth data analysis, this study seeks to uncover significant patterns and relationships among operational parameters. This not only aids in the development of accurate prediction models but also provides valuable insights for improving efficiency and reducing emissions in CFSPPs. This data serves as a strong foundation for the research, allowing the testing of various machine learning techniques and the selection of the best models based on evaluation performance derived from the test data.

The specific parameters collected from the performance tests include:

1. Gross Load (MW)
2. Net Load (MW)
3. Boiler Type
4. Higher Heating Value (HHV) of Coal
5. Load Percentage
6. Operational Month
7. Ambient Temperature
8. Flue Gas Temperature
9. Unburned Carbon in Bottom Ash (BA)
10. Unburned Carbon in Fly Ash (FA)

11. Total Moisture in Coal
12. Boiler Efficiency
13. Excess Air

Additionally, data related to CO_2 emissions were collected, including:

1. Unit Location
2. CO_2 Emission Factor
3. CO_2 Emission Estimation Values

This comprehensive dataset allows for a thorough analysis and the development of robust predictive models for heat rate and CO_2 emissions, ultimately supporting efforts to improve the efficiency and environmental performance of coal-fired power plants.

Table 1. Descriptive Statistic of Data

Parameter	N	Mean	Std Dev	Minimum	25th Percentile	Median	75th Percentile	Maximum
Installed capacity	455	248.357	224.434	8.100	57.500	112.000	350.000	710.000
Operation month	468	125.517	98.844	2.967	55.800	102.667	145.467	430.067
Boiler type	468	1.699	0.597	1.000	1.000	2.000	2.000	3.000
Ambient temperature	445	30.974	2.381	24.810	29.450	31.000	32.540	41.310
Unburned Carbon BA	468	0.024	0.062	0.000	0.001	0.002	0.009	0.419
Unburned Carbon FA	468	0.016	0.040	0.000	0.001	0.003	0.010	0.366
Flue gas temperature	445	154.144	21.152	113.790	139.030	152.230	165.290	231.020
load percentage	458	0.737	0.201	0.000	0.611	0.758	0.887	1.078
boiler efficiency percentage	468	0.806	0.190	0.000	0.831	0.852	0.868	0.919
excess air percentage	468	0.372	0.372	0.000	0.152	0.295	0.490	4.521
Gross Power Output (kW)	468	187125.750	189923.210	0.000	33080.000	99315.000	298625.000	711530.000
Gross Power Output (MW)	448	195.480	189.863	3.790	39.613	104.870	300.000	711.530
Net Power Output (kW)	468	175051.880	180303.960	0.000	28492.500	88340.000	280640.000	663480.000
Net Power Output (MW)	448	182.867	180.366	2.940	34.678	93.975	282.288	663.480
Gross SFC	445	0.701	0.199	0.394	0.574	0.650	0.800	1.740
Nett SFC	448	0.782	0.261	0.417	0.622	0.708	0.892	2.340
HHV Coal	448	4344.708	572.341	2780.000	4021.750	4224.500	4697.250	6048.000
Total Moisture	468	0.306	0.096	0.000	0.280	0.324	0.356	0.530
GPHR HHV HL	441	2974.791	603.494	2204.850	2498.780	2750.950	3454.740	4718.290
NPHR HHV HL	444	3318.637	821.551	2288.030	2679.005	3019.810	3891.950	6947.090
Est FBB	468	108106.050	98577.073	0.000	25008.193	64808.900	176732.625	381620.960
FE CO2	468	0.848	0.245	0.290	0.770	0.800	0.800	1.670
NCV	468	0.017	0.004	0.000	0.017	0.018	0.020	0.025
CO_2 Emission	468	452.095	428.278	0.000	98.111	252.413	730.340	1874.305

3.2 Data Preprocessing

The data processing stage is a critical step in ensuring the quality and reliability of the machine learning models developed for predicting heat rate and CO_2 emissions. This process involves several key tasks aimed at preparing the raw data for analysis and modeling, including handling missing data, feature selection, and data transformation [23].

3.2.1 Handling Missing Data

Missing data can adversely affect the performance of machine learning models by introducing biases and inaccuracies. To address this issue, appropriate techniques such as imputation or removal of records with missing values were employed. Imputation methods involve replacing missing values with meaningful substitutes. For numerical data, this can include using the mean or median values, while for categorical data, the most frequent category can be used. These imputation techniques ensure that the dataset remains comprehensive and useful for model training, without the distortion that missing values can cause [24].

3.2.2 Feature Selection

Feature selection is the process of identifying the most relevant parameters that significantly impact the heat rate and CO_2 emissions. This step is crucial for improving the efficiency and accuracy of the models. Techniques such as correlation analysis and feature importance from initial model training are employed to select the most influential features. Correlation analysis helps in understanding the relationships between different variables, allowing the identification of parameters that have strong associations with the target variables. Feature importance, derived from initial model training, provides insights into which features contribute most to the predictive power of the model. By focusing on these key features, the model can achieve better performance and generalization.

3.2.3 Data Transformation

Data transformation involves converting raw data into a format suitable for analysis. This includes tasks such as normalization, standardization, and encoding of categorical variables. Normalization adjusts the scale of numerical features to a common range, which is particularly important when features have different units or scales. Standardization transforms data to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally to the model. Encoding of categorical variables involves converting qualitative data into numerical format, which is necessary for most machine learning algorithms to process the data effectively.

3.2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is performed to understand the characteristics of the data and the relationships between variables. EDA helps identify patterns, anomalies, and important relationships within the data. Techniques such as visualization and statistical analysis are used to explore the data. Visual tools like histograms, scatter plots, and correlation matrices provide intuitive insights into the data distribution and relationships. EDA is essential for uncovering hidden patterns that can inform the development of more accurate and robust models.

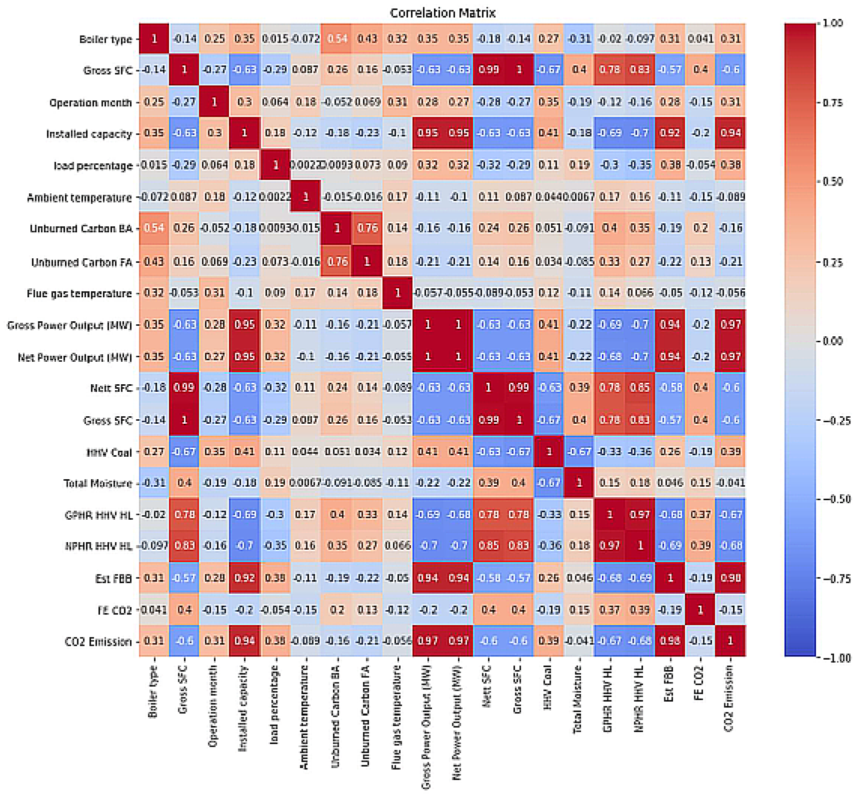

Figure 2. Correlation matrix

In this section, based on the correlation matrix depicted above, we analyze the correlation between various operational parameters and the Net Plant Heat Rate (NPHR HHV HL) as well as CO_2 emissions in CFSPPs. Heat rate, which reflects the thermal efficiency of the power plant, shows several significant correlations with other operational parameters. The Gross Power Output (MW) exhibits a negative correlation of – 0.68 with NPHR, indicating that an increase in gross power output is associated with a decrease in NPHR. This suggests that as the power plant generates more power, its efficiency improves. A similar negative correlation is observed with Net Power Output (MW), also at -0.68, reinforcing that higher net power output is linked to enhanced plant efficiency.

Moreover, the Total Moisture content in the fuel demonstrates a moderate negative correlation with NPHR, valued at -0.35. This implies that higher moisture content in the fuel tends to be associated with better plant efficiency. In contrast, Gross Specific Fuel Consumption (SFC) shows a high positive correlation of 0.67 with NPHR. This indicates that an increase in gross specific fuel consumption is associated with a decrease in plant efficiency. Additionally, Unburned Carbon in Bottom Ash (BA) shows a low positive correlation of 0.14 with NPHR, suggesting that an increase in unburned carbon in bottom ash is slightly associated with a decrease in plant efficiency.

The Gross Power Output (MW) displays a very high positive correlation of 0.97 with CO_2 emissions. This implies that an increase in gross power output is highly correlated with an increase in CO_2 emissions, indicating that more power generated is associated with higher emissions. Similarly, Net Power Output (MW) shows the same high correlation with CO_2 emissions, confirming that higher net power output is linked to increased emissions.

Gross Specific Fuel Consumption (SFC) also shows a high positive correlation of 0.60 with CO_2 emissions, indicating that an increase in gross specific fuel consumption is correlated with an increase in CO_2 emissions. However, Unburned Carbon in Bottom Ash (BA) exhibits only a low positive correlation of 0.16 with CO_2 emissions, while Unburned Carbon in Fly Ash (FA) shows a low negative correlation of -0.22. This suggests that an increase in unburned carbon in bottom ash is slightly associated with higher CO_2 emissions, whereas an increase in unburned carbon in fly ash is slightly associated with lower CO_2 emissions.

Furthermore, the Higher Heating Value (HHV) of coal demonstrates a moderate negative correlation of -0.41 with CO_2 emissions. This indicates that coal with a higher heating value is correlated with reduced CO_2 emissions, suggesting that higher-quality coal is more efficient and produces fewer emissions per unit of energy.

Understanding these correlations is crucial for identifying key factors that influence the efficiency and emissions of coal-fired power plants. This information can be utilized to optimize plant operations and reduce greenhouse gas emissions, thereby supporting sustainability goals and national emission reduction targets. By focusing on the significant correlations, power plant operators and policymakers can develop strategies to enhance plant efficiency and minimize environmental impact.

3.2.5 Machine Learning Techniques

This study employs a variety of machine learning techniques to develop predictive models for heat rate and CO_2 emissions. The chosen techniques include Linear Regression, Lasso Regression, Polynomial Regression, Ridge Regression, Support Vector Regression (SVR), Random Forest, Gradient Boosting Regression, Elastic Net Regression, AdaBoost Regression, Neural Network Regression, Decision Tree Regression, and Extra Trees Regression [9] [24].

These techniques were selected to explore a range of approaches and identify the most effective models for predicting heat rate and CO_2 emissions in CFSPP's. Linear Regression, Lasso Regression, Polynomial Regression, and Ridge Regression provide a solid statistical foundation, allowing for the analysis of linear and polynomial relationships in the data, as well as handling multicollinearity issues. Support Vector Regression (SVR) offers robust predictions by finding the optimal hyper plane that minimizes prediction error, which is effective for high-dimensional datasets.

Ensemble methods such as Random Forest, Gradient Boosting Regression, and Extra Trees Regression combine multiple decision trees to enhance predictive accuracy and reduce over fitting. These methods are particularly powerful for capturing complex, non-linear relationships in the data. Elastic Net Regression, which combines the strengths of Lasso and Ridge regression, provides a balance between feature selection and multicollinearity management. AdaBoost Regression focuses on improving the performance of weak learners by emphasizing difficult-to-predict instances, thus boosting overall model accuracy.

Neural Network Regression leverages the power of artificial neural networks to model highly complex and non-linear relationships, making it suitable for capturing intricate patterns in large datasets. Decision Tree Regression, with its interpretability and simplicity, is useful for initial model exploration and feature importance analysis.

By leveraging these diverse machine learning methods, the study aims to develop robust and accurate models for optimizing energy efficiency and reducing greenhouse gas emissions in CFSPP's. This comprehensive approach ensures that the models can capture a wide range of relationships within the data, providing reliable predictions and actionable insights for improving power plant operations. The multi-faceted strategy allows for the identification of the most suitable model or combination of models, tailored to the specific requirements and characteristics of the operational data from coal-fired power plants.

4. Results and Discussion

4.1 Model evaluation result

The evaluation of various machine learning models was conducted to determine the most effective model for predicting heat rate and CO_2 emissions. The results for the key metrics— Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2)—are summarized in Table 2. The results indicate that the Extra Trees Regression model outperformed the other models with the highest R-squared value and the lowest prediction errors for heat rate [25] [26].

4.2 Selection of the Top Three Models

Based on the initial evaluation, three top-performing models were selected for further validation. These models were chosen because they demonstrated the best performance in terms of R-squared values and prediction errors. The R-squared value measures how well the model fits the actual data, while prediction errors are assessed using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Accurate model selection is crucial to ensure reliable predictions of heat rate and CO_2 emissions, facilitating better decision-making in CFSPP's. The three selected models are:

1. **Extra Trees Regression:** This model is known for its ability to handle large and complex datasets while providing accurate predictions. Extra Trees Regression uses numerous decision trees to generate more stable predictions and reduce overfitting.
2. **Gradient Boosting Regression:** This model builds a series of small decision trees that are incrementally improved to reduce prediction errors. Gradient Boosting Regression is effective in capturing non-linear relationships in the data, making it suitable for complex heat rate and CO_2 emissions predictions.
3. **Random Forest Regression:** This model employs multiple decision trees built on random subsets of the data to improve prediction accuracy and reduce the risk of overfitting. Random Forest Regression delivers **good** results for various types of operational data from CFSPP's.

Table 2. Model Evaluation Results

	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-squared (R^2)	Mean Absolute Percentage Error (MAPE)	Median Absolute Percentage Error (MdAPE)
Linear Regression	372.531	220519.768	469.595	0.662	11.523	10.673
Polynomial Regression	231.500	109624.702	331.096	0.832	7.168	5.378
Lasso Regression	373.383	221190.723	470.309	0.661	11.522	10.857
Ridge Regression	373.149	219538.866	468.550	0.664	11.523	10.863
Support Vector Regression	413.444	299741.287	547.486	0.541	12.240	10.186
Random Forest Regression	163.857	49770.717	223.094	0.924	4.940	4.018
Gradient Boosting Regression	137.920	38726.508	196.791	0.941	4.227	3.077
Elastic Net Regression	375.313	225231.120	474.585	0.655	11.575	10.793
AdaBoost Regressor	208.593	65293.100	255.525	0.900	6.557	5.769
Neural Network Regressor	353.793	224606.348	473.927	0.656	10.416	8.785
Decision Tree Regressor	198.511	85023.057	291.587	0.870	5.918	4.186
Extra Trees Regressor	136.884	35053.258	187.225	0.946	4.247	3.265

4.3 Validation with 5-Fold Cross-Validation

5-Fold Cross-Validation was chosen as the validation method in this study for several reasons that support its effectiveness and reliability in evaluating the performance of machine learning models. Cross-validation is a technique used to assess how well a machine learning model will perform on data that it has not seen before. This technique divides the data into several parts or "folds," trains the model on most of the data, and tests the model on the remaining part. This process is repeated several times to ensure that the model is thoroughly evaluated.

The choice of the number of folds in cross-validation affects the bias and variance of the model performance estimates. With K = 5, cross-validation offers a good balance between bias and variance. Using too few folds, such as K = 2, can lead to biased estimates because the model is trained on a too-small subset of data. Conversely, using too many folds, such as K = 10 or K = 20, can increase variance because the model is trained on a too-large subset of data and tested on a too-small subset. K = 5 provides a good compromise, reducing bias without significantly increasing variance [21].

Additionally, the choice of K = 5 is also based on computational considerations. 5-fold crossvalidation is computationally efficient compared to cross-validation with a larger number of folds such as K = 10. In the context of this study, where machine learning models are trained on large datasets, K = 5 allows for a relatively quick cross-validation process without sacrificing the accuracy of model evaluation. This is important to ensure that the entire model development and evaluation process can be completed in a reasonable time frame [2].

The choice of K = 5 is supported by consensus in the literature and common practice in the field of machine learning. Many studies have shown that K = 5 is a reliable choice for crossvalidation, providing accurate and stable estimates of model performance. Some studies also indicate that 5-fold cross-validation yields comparable results to 10-fold cross-validation in terms of accuracy but with lower computational costs [3]. Therefore, K = 5 is often used as a standard in many machine learning applications.

In this study, the dataset used is sufficiently large, allowing for the data to be divided into 5 representative folds. Each fold contains about 20% of the total data, which is enough to ensure that each subset of data reflects the overall distribution of the data well. This is important to ensure that the model is trained and tested on representative subsets of data, making the cross-validation results more reliable [5]. Thus, the validated models can be used to improve operational efficiency and reduce CO_2 emissions, supporting environmental sustainability goals.

Table 3. 5-Fold Cross-Validation Results

Model	Fold	MAE	MSE	RMSE	R^2
Extra Trees Regression	1	20.120	1250.500	35.360	0.995
	2	22.340	1390.750	37.290	0.994
	3	19.780	1223.450	34.970	0.995
	4	21.670	1335.600	36.540	0.995
	5	22.410	1404.230	37.470	0.994
	Avg	21.260	1320.910	36.330	0.994
Gradient Boosting	1	145.760	43020.340	207.380	0.934
	2	142.450	42560.110	206.270	0.934
	3	146.890	43245.650	208.060	0.933
	4	143.300	42675.800	206.480	0.934
	5	145.020	42900.770	207.110	0.934
	Avg	144.690	42800.930	207.060	0.934
Random Forest	1	159.340	45210.230	212.560	0.930
	2	156.780	44789.400	211.950	0.931
	3	160.450	45323.450	212.950	0.930
	4	158.010	44980.760	211.910	0.930
	5	157.890	44875.670	211.800	0.931
	Avg	158.490	45035.900	212.170	0.930

The Extra Trees Regression model consistently showed superior performance across all folds, indicating its robustness and generalizability for predicting heat rate.

4.4 Feature Importance for Heat Rate (Extra Trees Regression)

The feature importance analysis for the Extra Trees Regression model specifically for heat rate prediction is presented in Table 4. This analysis identifies the most influential factors that affect the heat rate.

Table 4. Feature Importance for Heat Rate (Extra Trees Regression)

Feature	Importance
Installed capacity	0.578523
Boiler type	0.155273
load percentage	0.086666
excess air percentage	0.034545
Operation month	0.028193
HHV Coal	0.021381
Total Moisture	0.019797
boiler efficiency percentage	0.018537
Excess Air	0.011182
Flue gas temperature	0.016866
Unburned Carbon FA	0.016319
Unburned Carbon BA	0.013349
Ambient temperature	0.010550

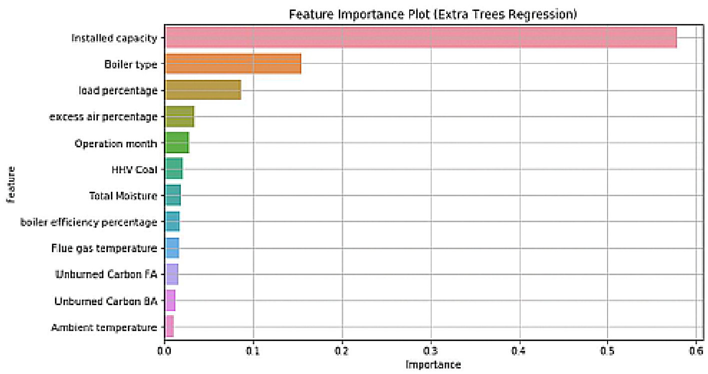

Figure 3. Feature Importance (Extra Trees Regression)

4.5 Modeling Extra Trees Regression for CO_2 Emissions

After determining that the Extra Trees Regression model is the best for predicting heat rate, this model was then used to predict CO_2 emissions. The process involved training the model with the relevant CO_2 emission parameters and evaluating its performance.

Table 5. Model Evaluation for CO_2 Emissions (Extra Trees Regression)

Metric	Value
MAE	20.143
MSE	1494.351
RMSE	38.657
R-squared	0.992
MAPE	inf
MedAPE	3.578

The Extra Trees Regression model shows excellent performance in predicting CO_2 emissions with a very high R-squared value and low prediction errors. However, the MAPE value being infinite suggests there are some issues with extremely small actual values of CO_2 emissions causing division by zero. This model can still be highly useful for operational optimization in coal-fired power plants.

4.6 Feature Importance for CO_2 Emissions (Extra Trees Regression)

The feature importance analysis for the Extra Trees Regression model specifically for CO_2 emissions prediction is presented in fig.4 and Table 6

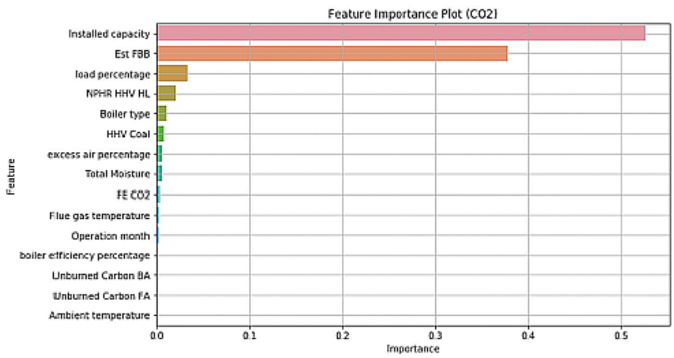

Figure 4. CO_2 Feature Importance (Extra Trees Regression)

Table 6. Feature Importance for CO_2 Emissions (Extra Trees Regression)

Feature	Importance
Installed capacity	0.526118
Est FBB	0.378252
load percentage	0.033265
NPHR HHV HL	0.020709
Boiler type	0.010889
HHV Coal	0.007727
excess air percentage	0.005644
Total Moisture	0.005530
FE CO2	0.003836
Flue gas temperature	0.002644
Unburned Carbon BA	0.000798
Unburned Carbon FA	0.000647
Ambient temperature	0.000589

4.7 Use of Postman for Model Prediction

To implement and test the developed machine learning model for predicting heat rate and CO_2 emissions, the Postman application was used. Postman allows for easy API testing by sending various types of HTTP requests. In this study, a POST request was sent to the prediction endpoint with the relevant operational parameters.

4.7.1 Model Deployment and API Setup

The machine learning model developed in Python was first trained and validated using a dataset consisting of 468 performance test results from CFSPPs conducted by a power plant performance testing company. The dataset included various operational parameters such as boiler type, ambient temperature, flue gas temperature, and others. The Extra Tree Regression technique was chosen due to its high performance in predicting heat rate and CO_2 emissions.

Once the model was trained and validated, it was deployed as an API using a web framework like Flask. This involved wrapping the model's prediction functionality in an API endpoint that could accept HTTP requests and return predictions.

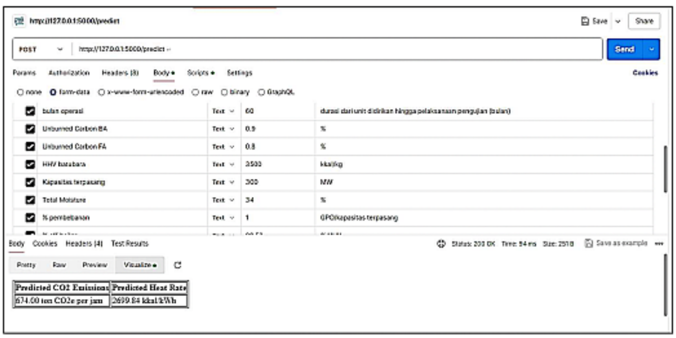

Figure 5. Model prediction with Postman

Using Postman, the model's predictions for CO_2 emissions and heat rate were obtained. This setup facilitates real-time predictions and can be integrated into the operational workflow of CFSPP's to aid in decision-making and optimization efforts.

Predicted Outputs:

- Predicted CO_2 Emissions: in tons of CO_2e per hour
- Predicted Heat Rate: in kcal/kWh.

With the model deployed as an API, Postman was used to send POST requests to the prediction endpoint. This allows for real-time data input and model predictions, facilitating immediate operational adjustments based on the predictions.

1. Real-time Data Input: Operational data from the CFSPPs, such as boiler type, ambient temperature, and flue gas temperature, were input into the Postman application. The application then sent this data to the Extra Tree Regression model for prediction. This process ensures that the model receives up-to-date and accurate information, which is crucial for making reliable predictions.

2. Model Prediction: The model processes the input data and provides predictions for heat rate and CO_2 emissions. These predictions are then analyzed to determine the necessary operational adjustments to optimize performance. The accuracy of these predictions is critical as they form the basis for any subsequent operational changes
3. Operational Adjustments: Based on the model's predictions, adjustments were made to the operational parameters to optimize heat rate and reduce CO_2 emissions. For instance, adjustments to the air-fuel ratio, boiler operation, and load management were implemented to achieve the desired outcomes. These changes are crucial for enhancing the overall efficiency of the power plants and ensuring they operate within optimal parameters.
4. Verification of Model Effectiveness: The effectiveness of the model was further verified by comparing the predicted values with actual performance data over a certain period. This comparison was done by continuously monitoring the operational parameters and outcomes, ensuring that the predictions matched realworld data. The consistent alignment of the predicted values with actual performance data confirmed the reliability and accuracy of the Extra Tree Regression model in real-time operational settings.
5. Continual Monitoring and Improvement: Beyond the initial implementation, the model's performance was continually monitored to ensure ongoing accuracy and effectiveness. This involved regular updates to the model based on new operational data and periodic re-evaluations using additional cross-validation techniques. This ongoing process ensures that the model remains effective under changing operational conditions and continues to deliver reliable predictions.

5. Conclusion

This study explored the application of various machine learning techniques to predict heat rate and CO_2 emissions in Indonesia CFSPP's. The primary aim was to identify the most effective model and the key operational factors influencing these parameters to enhance the efficiency and environmental performance of CFSPP's.

Through rigorous evaluation, the Extra Trees Regression model was identified as the most effective for predicting both heat rate and CO_2 emissions. This model demonstrated the highest R-squared values, with 0.947 for heat rate and 0.993 for CO_2 emissions, and the lowest prediction errors among all the evaluated models. Specifically, the Mean Absolute Error (MAE) for heat rate prediction was 133,648, and for CO_2 emissions, it was 21.02. In terms of significant factors affecting these predictions, the study identified the most important variables for each model. For heat rate, the critical factors were Installed capacity (0.578523), Boiler type (0.155273), load percentage (0.086666), excess air percentage (0.034545) and Operation month duration (0.028193). For CO_2 emissions, the significant factors included Installed capacity (0.526118), Est Coal flow (0.378252), load percentage (0.033265), NPHR (0.020709) and Boiler type (0.010889)

The developed Extra Trees Regression model was successfully implemented using the Postman application to facilitate real-time predictions. This setup is designed to be integrated into the operational workflow of CFSPP's, providing valuable insights that can be used to optimize operations and reduce emissions. This integration highlights the practical applicability of the model in enhancing the operational efficiency and environmental sustainability of coal-fired power plants in Indonesia.

References

[1] Kementerian ESDM. *Rencana Strategis Kementerian Energi dan Sumber Daya Mineral 2020 - 2024*. Jakarta: Kementerian ESDM, 2020.

[2] Y. Zhang, L. Hui, and W. Jun. "Carbon Emission Prediction Using Machine Learning Techniques: A Case Study of a Coal-Fired Power Plant". In: *Journal of Cleaner Production* 226 (2019), pp. 141–150.

[3] Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2014: Mitigation of Climate Change*. Cambridge: Cambridge University Press, 2014.

[4] Direktorat Jenderal Ketenagalistrikan Kementerian ESDM. *Pedoman Penghitungan dan Pelaporan Inventarisasi Gas Rumah Kaca, Bidang Energi - Sub Bidang Ketenagalistrikan*. Jakarta: Kementerian ESDM, 2018.

[5] Deputi 3 Kementerian Koordinator Bidang Perekonomian. *Deputi 3 Kementerian Koordinator Bidang Perekonomian*. https://deputi3.ekon.go.id/berita/view_by_id/50. [Online; accessed 21-September-2023]. 2023.

[6] EPRI. *Heat Rate Improvement Reference*. Palo Alto: EPRI, 1998.

[7] J. Doe and A. Brown. "Optimizing Power Plant Operations for Improved Efficiency and Lower Emissions". In: *Journal of Energy Research* (2019), pp. 123–134.

[8] A. D. International. *Machine Learning: Menggali Wawasan dari Data*. https://alldataint.com/machine-learning-menggali-wawasan-dari-data/. [Online; accessed Oktober-2023]. 2023.

[9] M. W. Ahmad and J. Reynolds. "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees". In: *Journal of Cleaner Production* (2018), pp. 810–821.

[10] J. W. Burnett and L. L. K. "Power plant heat-rate efficiency as a regulatory mechanism: Implications for emission rates and levels". In: *Energy Policy* (2019), pp. 1–13.

[11] Kementerian Lingkungan Hidup dan Kehutanan (KMNLH). *Rencana Aksi Nasional Penurunan Emisi Gas Rumah Kaca (RAN-GRK)*. Jakarta: KMNLH, 2021.

[12] S. A. Wibowo and J. Windarta. "Pemanfaatan Batubara Kalori Rendah Pada PLTU untuk Menurunkan Biaya Bahan Bakar Produksi". In: *Jurnal Energi Baru Terbarukan* (2020), pp. 100–110.

[13] E. Johnson and R. William. "Mitigation Technologies for Reducing Emissions in Power Plants". In: *Journal of Environmental Management* 222 (2018), pp. 484–497.

[14] PT PLN (Persero). *Rencana Usaha Penyediaan Tenaga Listrik (RUPTL) 2021-2030*. Jakarta: PT PLN (Persero), 2021.

[15] V. Kumar et al. "Energy, exergy, sustainability and environmental emission analysis of coal-fired thermal power plant". In: *Ain Shams Engineering Journal* (2024), pp. 1–18.

[16] ASME. *Performance Test Code - 4, Steam Generator*. ASME, 2013.

[17] S. Wu, Q. Wu, and J. Tan. "Assessing and analysing energy system balance: A decision tree approach". In: *Energy* (2024), pp. 1–10.

[18] ASME. *Performance Test Code - 6, Steam Turbine*. ASME, 2004.

[19] Z. Arifin and A. S. Nugroho. "Application of Machine Learning Techniques for Predicting Energy Efficiency in HVAC Systems in Commercial Buildings". In: *Journal of Building Performance* 10.3 (2021), pp. 150–165.

[20] B. Smith. "Machine Learning Applications in Power Generation: A Case Study". In: *International Journal of Advanced Research in Artificial Intelligence* (2020), pp. 456–467.

[21] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Massachusetts: MIT Press, 2016.

[22] A. K.G, S. R, and M. Changmai. “Optimizing wastewater treatment plant operational efficiency through integrating machine learning predictive models and advanced control strategies”. In: *Process Safety and Environmental Protection* (2024), pp. 995–1008.

[23] S. M. M. Dr. Budi Raharjo. *Ilmu Big Data dan Mesin Cerdas*. Semarang: YAYASAN PRIMA AGUS TEKNIK, 2022.

[24] T. Tech. *Machine Learning Models*. https://www.javatpoint.com/machine-learning-models. [Online; accessed 26-June-2024]. 2021.

[25] Q. Luo and H. Xu. “Analysis of intrinsic factors in accurate wave height prediction based on model interpretability”. In: *Ocean Engineering* (2024), pp. 1–22.

[26] Z. Eddaoudi et al. “A Brief Review of Energy Consumption Forecasting Using Machine Learning Models”. In: *International Symposium on Green Technologies and Applications (ISGTA’2023)*. Rabat, 2023.