Received 22 February 2024, accepted 13 March 2024, date of publication 22 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380809

APPLIED RESEARCH

A Hybrid Predictive Model as an Emission Reduction Strategy Based on Power Plants' Fuel Consumption Activity

NENG AYU HERAWATI[1], ASYRAF ATTHARIQ PUTRA GARY[2], ERNA HIKMAWATI[3], AND KRIDANTO SURENDRO[1], (Member, IEEE)
[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40116, Indonesia
[2]Independent Researcher, Jakarta 13130, Indonesia
[3]School of Applied Science, Telkom University, Bandung 40257, Indonesia

Corresponding author: Neng Ayu Herawati (nengayu@itb.ac.id)

ABSTRACT The Indonesian government has announced the initiation of carbon trading in 2023, commencing with coal-fired power plants (PLTUs). Despite this, certain PLTUs in Indonesia, including one examined in this research, still rely on manual identification of greenhouse gas (GHG) emissions and need more data to support forthcoming carbon trading endeavors. To address this gap, this study focuses on developing a predictive model using The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology to forecast carbon emissions, coal fuel consumption, and gross electricity. The predictive model integrates time series models to forecast each variable, followed by a regression model for carbon emission prediction. Evaluation metrics, including MAE, RMSE, and MAPE, are utilized, and model training employs five machine learning models: Linear Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and LightGBM. Results reveal that for PLTU Units 3 and 7, the Linear Regression model performs optimally in time series modeling, whereas for PLTU Unit 8, Random Forest Regression is most effective. Across all units, Linear Regression emerges as the superior model in regression modeling. This study equips PLTUs with predictive insights into carbon emissions, facilitating strategic planning for emission reduction. By considering predicted outcomes of coal fuel consumption and gross electricity alongside carbon emission forecasts, PLTUs can comprehensively assess environmental and financial impacts, guiding effective mitigation strategies.

INDEX TERMS Coal-fired power plants, emissions prediction, greenhouse gas emissions, machine learning.

I. INTRODUCTION

Global warming is a phenomenon of increasing global temperature on Earth due to the effect of Greenhouse Gases (GHG), thus increasing environmental problems around the world [1], [2], [3]. One source of GHGs is emissions from burning coal for power generation. More than 40% of total carbon dioxide (CO2) emissions associated with energy are attributable to the process of burning fossil fuels for electricity generation purposes [4], [5]. The Indonesian government is dedicated to achieving a 29% reduction in greenhouse gas (GHG) emissions by 2030, amounting to 834 million tons of CO2e. Emphasizing the energy sector as a key focus, a target of 314 million tons of CO2e reduction has been set [6]. The government is actively curbing GHG emissions from coal-fired power plants through a carbon trading mechanism, slated to commence in 2023, per the Ministry of Energy and Mineral Resources [7]. Consequently, each power generation unit is mandated to calculate GHG emissions, and various agencies must optimize efforts to minimize carbon emissions according to the Carbon Economic Value implementation mechanism. This necessitates power plants to operate electricity generation efficiently in compliance with relevant legal provisions [8], [9].

The case study in this research takes one of Indonesia's private coal-fired power plants (PLTU). This power plant has

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

three operating units: Unit 3, Unit 7, and Unit 8. The power plant has identified GHG emissions based on activity data per fuel type but has yet to develop a system to predict carbon emissions. Awareness of the importance of information on future emission levels is one of the crucial steps in planning to reduce GHG emissions [10]. Based on this, emission prediction can be used to implement emission reduction strategies effectively [11]. At the United Nations COP26 Conference on Climate Change in 2021, opportunities to reduce GHG emissions were examined in 28 different axes as part of efforts to prevent global warming. One of the axes examined is the influence of machine learning on GHG emission prediction in assisting informative decision-making to reduce and manage GHG emissions [12]. Based on this, carbon emission prediction analysis can be implemented using a variety of methods, including machine learning-based approaches [13], [14], [15], [16].

Currently, several studies have implemented machine-learning approaches to carbon emission prediction. The majority of existing research only uses global or country carbon emissions data for the process of developing predictive models [15], [17], [18], [19]. Furthermore, some other studies only use internal data, so they do not consider various external factors that may relate to emissions prediction [16]. Getting real-time data on fuel combustion activities at a coal-fired power plant that can be processed into carbon emission data in model development is a big challenge. Based on this, this study will develop a carbon emission prediction model using historical data on fuel combustion activities by utilizing features related to emission prediction, one of which is the use of coal feeder equipment at PLTU to identify coal fuel consumption and electrical energy generated in real-time with granularity per day.

This research will focus on developing a hybrid time series and regression models. This research will develop a prediction model for each feature related to carbon emission prediction (such as coal fuel consumption, the electrical energy produced, weather data, and others) using a time series model to make a prediction feature that will be used as input to the regression model to predict emissions utilizing the relationship of independent and dependent variables. Thus, the resulting model will combine the time series model as a prediction feature whose results will be integrated with the emission prediction model using a regression model so that it can produce emission predictions.

With accurate predictions of carbon emissions, PLTU would be able to estimate future carbon trading decisions. Similarly, with insights into predicted fuel consumption and gross electricity output, PLTU can easily determine fuel consumption estimates based on emission predictions and carbon trading decisions. Thus, integrating predictive models for carbon emissions, fuel consumption, and gross electricity becomes paramount for informed decision-making in carbon trading and emission reduction strategies. These models enable PLTU to anticipate environmental impacts and facilitate financial planning and risk management in future trading endeavors. By leveraging advanced data mining methodologies such as The Cross Industry Standard Process for Data Mining (CRISP-DM) and employing various machine learning algorithms like Linear Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and LightGBM, this research aims to develop robust predictive models tailored to the complexities of PLTU operations. Synthesis of time series and regression models allows the study to provide PLTU with actionable insights into emission trends, fuel consumption patterns, and electricity generation dynamics, empowering proactive carbon management strategies and sustainable operational practices.

II. RELATED WORKS

This section will discuss all related works that are used as the primary reference and used in developing solutions.

A. CALCULATION OF GHG EMISSIONS IN THE POWER GENERATION SECTOR

The types of GHG emissions calculated by fossil-fuel power generation units and biomass-based fuels are carbon dioxide gas (CO2), methane gas (CH4), and nitrous oxide gas (N2O). In total, three types of gases need to be calculated so that it is necessary to convert non-CO2 GHG emissions into carbon dioxide equivalent (CO2e) using the Global Warming Potential (GWP) value contained in the Second Assessment Report of IPCC (2nd AR of IPCC) [8]. The calculation of CO2 emissions uses a calculation and measurement-based approach. Each power plant can choose various calculation methods according to the data availability in the field. The Ministry of Energy and Mineral Resources has provided guidelines for calculating emissions using four methods. The emission calculation method uses activity data per type of fuel at each coal-fired power plant (Method 1, Method 2, Method 3) and emission calculation using the Continuous Emissions Monitoring System (CEMS) or Method 4 [8]. This research will use Method 1 in calculating CO2 emissions. The formula for calculating CO2 emissions in Method 1 can be seen in Formula 1.

$$\mathrm{ECO2} = \mathrm{DA} \times \mathrm{FE} \tag{1}$$

Activity data is fuel consumption data that has been converted into units of energy. The formula for converting coal fuel consumption data from units of mass (tons) to units of energy (TJ) in calculating activity data in calculating CO2 emissions Method 1 can be seen in Formula 2.

$$\mathrm{DABB} = \mathrm{FBB} \times \mathrm{NCV} \times 10^{-3} \tag{2}$$

Description:

DABB :Coal Activity Data (TJ)
FBB :Coal consumption (ton)
NCV :The Net Calorific Value (NCV) of coal (TJ/Gg), the NCV value of specific coal-generating units, or default national

The coal-fired power plant uses diesel fuel in addition to it. Diesel fuel is included in the category of Fuel Oil (BBM). The formula for converting fuel consumption data from kilo liters to units of energy (TJ) in calculating activity data in calculating CO2 emissions Method 1 can be seen in Formula 3.

$$\mathrm{DABBM} = \mathrm{FBBM} \times \rho \times \mathrm{NCV} \times 10^{-6} \tag{3}$$

Description:

DABBM : Fuel oil Activity Data (TJ)
FBBM : Fuel oil consumption (kiloliters)
NCV : The Net Calorific Value (NCV) of fuel (TJ/GgBBM), the NCV value of specific fuel oil-generating units, or default national
ρ : The specific gravity of fuel (kgBBM/m^3)

This power plant also uses biodiesel as an additional fuel. Biodiesel fuel falls into the category of liquid biomass. The formula for converting fluid biomass consumption data from kilo liters to units of energy (TJ) in calculating activity data in calculating CO2 emissions Method 1 can be seen in Formula 4.

$$\mathrm{DABM} = \mathrm{FBM} \times \mathrm{NCV} \times \rho \tag{4}$$

Description:

DABM :Liquid biomass activity data (TJ)
FBM :Consumption of liquid biomass (kilo liters)
NCV :The Net Calorific Value (NCV) of liquid biomass (TJ/Gg), the specific NCV value of generating unit or default national
ρ :The specific gravity of liquid biomass (kg/m^3)

In addition to calculating CO2 emissions, each plant must calculate CH4 and N2O emissions into carbon dioxide equivalent (CO2e) using the GWP value. CH4 and N2O emissions from fuel combustion can be calculated using Method 1. The following is a formula for calculating CH4 and N2O emissions using Method 1 for coal fuel, fuel, and liquid biomass. The formula for calculating CH4 and N2O emissions of coal fuel can be seen in Formula 5.

$$\mathrm{ECH4\ and\ N2O\ BB} = (\mathrm{FBB} \times \mathrm{NCV}) \times \mathrm{FE} \times 10^{-6} \tag{5}$$

Description:

ECH4, N2O :Emission Factor (kg CH4/TJ or kg N2O/TJ)
FBB :Coal consumption (ton)
NCV :The Net Calorific Value (NCV) of coal (weighted average, TJ/GG)

The formula for calculating CH4 and N2O emissions of Fuel Oil (BBM) can be seen in Formula 6.

$$\mathrm{ECH4\ and\ N2O\ BBM} = (\mathrm{FBBM} \times \mathrm{NCV} \times \rho) \times \mathrm{FE} \times 10^{-9} \tag{6}$$

Description:

ECH4, N2O :Total CH4 and N2O emissions (tonnes)
FE :Emission Factor (kg CH4/TJ or kg N2O/TJ)
FBBM :Fuel oil consumption (kilo liters)
NCV :The Net Calorific Value (NCV) of fuel (weighted average, TJ/GG)
ρ :The specific gravity of fuel oil (kgBBM/m^3)

The formula for calculating CH4 and N2O emissions of liquid biomass fuel can be seen in Formula 7.

$$\mathrm{ECH4\ and\ N2O\ BM} = \mathrm{FBM} \times \rho \times \mathrm{NCV} \times \mathrm{FE} \times 10^{-6} \tag{7}$$

Description:

ECH4, N2O :Total CH4 and N2O emissions (tonnes)
FE :Emission Factor (ton CH4/TJ or ton N2O/TJ)
FBM :Consumption of liquid biomass (kilo liters)
NCV :The Net Calorific Value (NCV) of liquid biomass (weighted average, TJ/GG)
ρ :The specific gravity of liquid biomass (kg/m^3)

B. EMISSION LIMIT REGULATION IN INDONESIA

In 2020, Indonesia initiated a carbon emissions trading trial targeting coal-fired power plants. According to information from the Directorate General of Electricity, 32 coal-fired power plants participated in the carbon emission trading program. Among these, 14 coal-fired power plants operated as buyers, constituting 44% of the total, while 18 coal-fired power plants functioned as sellers, making up the remaining 56%. The cumulative volume of carbon transactions during this period reached 42,455.42 tons of CO2e, with an average unit price of $2 per ton of CO2 [20].

A prerequisite for carbon trading frameworks involves implementing a cap or an upper threshold for greenhouse gas (GHG) emissions. The upper limit for emissions constitutes a technical accord established by the Minister of Energy and Mineral Resources, delineating the permissible emission levels within a specified timeframe [9]. Each unit's upper emission limit is determined at this power plant per the Minister of Energy and Mineral Resources Decree [21]. The 2023 Upper Emission Limit values for the power plant are detailed in Table 1, elucidating distinct cap values for three operational coal-fired power plant units based on their respective installed capacity levels.

C. CARBON EMISSION PREDICTION MODELS USING MACHINE LEARNING

This section contains a literature review comprising a series of studies relevant to the research, serving as a reference source. A thorough literature review can enhance the comprehension of the upcoming research. The pertinent studies are outlined in Table 2. Based on Table 2, there are fundamental differences in this study. This research will develop a combination of time series models to predict each variable used as a feature to predict emissions. Then, the prediction results are used as input to regression models in predicting carbon emissions. Furthermore, the variables in this study are quite varied, including variables of coal fuel consumption, gross electricity, CO2 emissions, CO2e emissions, and external data, namely weather data. Then, the data used using specific data from the power plant uses data with granularity per day to generate more data to improve model performance.

TABLE 1. Power plant's PTBAE value in 2023.

Unit Name	Capacity (MW)	Stamp (tons CO2e)
Unit 3	815	5.384.763,31
Unit 7	615	3.949.602,30
Unit 8	815	3.891.324,51

III. PROPOSED METHOD

This section explains in detail the development of the model to be built. Model development in this study is one of the developments from previous research [13]. The methodology used can be seen in Figure 1.

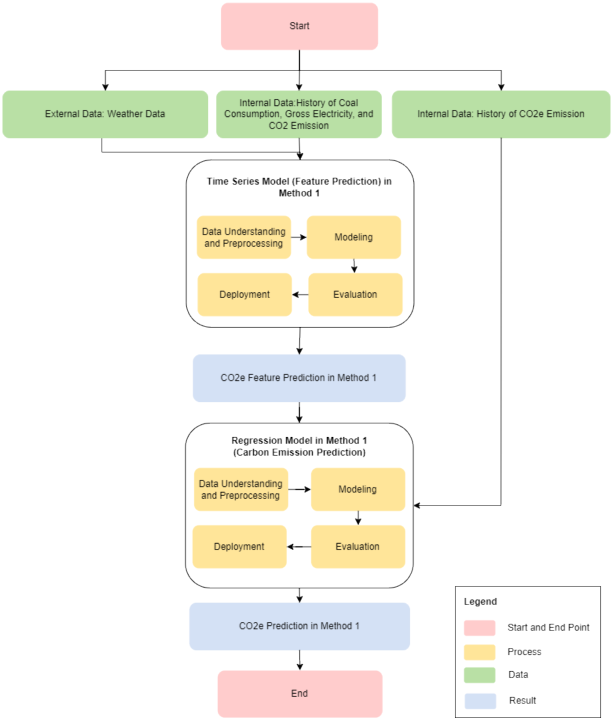

Start
External Data: Weather Data
Internal Data:History of Coal Consumption, Gross Electricity, and CO2 Emission
Internal Data: History of CO2e Emission
Time Series Model (Feature Prediction) in Method 1
Data Understanding and Preprocessing
Modeling
Deployment
Evaluation
CO2e Feature Prediction in Method 1
Regression Model in Method 1 (Carbon Emission Prediction)
Data Understanding and Preprocessing
Modeling
Deployment
Evaluation
CO2e Prediction in Method 1
End
Legend
Start and End Point
Process
Data
Result

FIGURE 1. Methodology.

Based on Figure 1, the initial part of the model development stages employs The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology due to its comprehensive framework that systematically guides the planning and execution of predictive modeling. CRISP-DM stands out for its structured approach to understanding both business objectives and the intricacies of the data involved. This methodology is pivotal in ensuring the development process is aligned with the stakeholders' goals and robust enough to address the underlying issues effectively. This structured approach is crucial for successfully predicting carbon emissions using machine learning, as it ensures that the models are developed with a clear understanding of the objectives, data, and potential challenges.

The model development phase in CRISP-DM begins with data understanding, followed by data preparation, modeling, evaluation, and deployment. The input data for predictive modeling includes both internal and external datasets. Predictive modeling involves using time series models to predict each feature within the regression model. Subsequently, the regression model generates predictions of carbon emissions. This sequential approach ensures a comprehensive analysis and data utilization, allowing for accurate predictions and effective decision-making strategies for reducing carbon emissions.

A. DATA UNDERSTANDING

Data collection in this study uses two primary data, namely internal data, which is activity data generated from the fuel combustion process at the power plant, and external data, namely weather data, to see the correlation with internal data.

1) INTERNAL DATA

The dataset utilized in this study originates from a private PLTU in Indonesia, covering the period from January 1, 2021, to July 31, 2023. It provides a comprehensive overview of emissions for Units 3, 7, and 8, utilizing Method 1. Specifically, the data is sourced from the coal feeder equipment operational at the PLTU, capturing information daily over a 24-hour cycle. This equipment is responsible for precisely metering and delivering coal to the power generation process. By capturing data from the coal feeder equipment daily, we gain a granular understanding of coal consumption patterns, which directly influence emissions within the PLTU. This data source offers valuable insights into the operational dynamics of the power plant, furnishing essential variables necessary for analyzing emission trends and guiding the development of predictive models aimed at emission reduction strategies. The dataset encompasses a mix of independent and dependent variables critical for understanding and analyzing the emission dynamics within these units.

a: INDEPENDENT VARIABLES

1. Coal Fuel Consumption Data: Represents the quantity of coal used within a specific time frame. This metric is pivotal as it directly influences the electricity generated and the volume of emissions produced. High coal consumption typically correlates with increased power output and higher emission levels.
2. Gross Electricity or Electrical Energy Results from the Combustion Process: Quantifies the total electricity generated from burning coal. This variable is essential for assessing the efficiency and environmental impact of the power generation process. It bridges fuel consumption and emission production, highlighting energy conversion effectiveness from coal to electricity.

TABLE 2. Literature review.

Research Title	Data Used	Scope	Method	Metrics	Advantages	Limitation
A Hybrid Model with Applying Machine Learning Algorithms and Optimization Model to Forecast Greenhouse Gas Emissions with Energy Market Data [3].	Global dataset (Statistical Center and the Ministry of Energy of Iran).	Annual	Time series models and mathematical models.	RMSE, NRMSE, MAPE, MAE, and RAE.	Make long-term predictions by combining machine learning models with mathematical models.	Single variable, only using historical emission data to make emission predictions, small dataset (15 years or 15 rows) because of granularity per year.
Prediction Model: CO2 Emission Using Machine Learning [15].	Global dataset (World Bank).	Annual	Time series model.	RMSE.	Resulting in a low RMSE value.	Single variable, only using history emission data to predict emissions, only conducting experiments using linear regression model, small dataset (54 years or 54 rows) due to granularity per year.
Carbon Emission Prediction of Thermal Power Plants Based on Machine Learning Techniques [16].	Three power plants are located in Northwest China.	Monthly	Regression model.	RMSE	Using seven independent variables that vary.	The results of emission predictions depend on independent variables, so they cannot make long-term predictions because there are no predictions available for each independent variable, and there are small datasets (31 months or 31 rows) due to granularity per month.
Carbon Emission Prediction with Macroeconomic Variables and Machine Learning [17].	Global dataset (National Bureau of Statistics of China).	Annual	Regression model.	RMSE, MAE, and MAPE.	Using 44 independent variables that vary.	The results of emission predictions depend on independent variables, so they cannot make long-term predictions because there are no predictions available for each independent variable, and there are small datasets (9 years or 9 rows) due to granularity per year.
Comparison of Carbon Emission Forecasting in Guangdong Province Based on Multiple Machine Learning Models [18].	Global dataset (National Bureau of Statistics and the Guangdong Provincial Statistical Yearbook)	Annual	Time series model.	RMSE, MAE, MAPE, R2.	Comparing seven machine learning models.	Single variable, only using historical emission data to predict emissions, small dataset (14 years or 14 rows) because of granularity per year.

3. CO2 Emission Data: Tracks the volume of carbon dioxide released during the coal combustion process. CO2 emissions are a primary concern for environmental and climate change studies, as they contribute significantly to the greenhouse gas effect. Monitoring these emissions is crucial for evaluating the environmental footprint of power generation activities.

b: DEPENDENT VARIABLE

CO2e Emission Data (CO2, N2O, and CH4 emissions): Provides a comprehensive measure of greenhouse gas emissions, encapsulating not just CO2 but also nitrous oxide (N2O) and methane (CH4). The CO2e (carbon dioxide equivalent) metric aggregates the impact of all these gases into a single measure, reflecting the total greenhouse gas emissions in terms of CO2's equivalent implications. This variable is dependent on the independent variables, as changes in coal consumption, electricity generation, and direct CO2 emissions influence the overall CO2e emissions. Including N2O and CH4, alongside CO2, in the CO2e calculation is vital for capturing the full spectrum of greenhouse gas emissions, given their potent global warming potential.

2) EXTERNAL DATA

Integrating external weather data from Meteum AI offers a distinctive insight into the environmental factors impacting power plant operations and emissions [22]. This dataset, which was collected daily from January 1, 2021, to July 31, 2023, encompasses several critical variables. These variables are employed as independent variables in predicting the dependent variable, namely CO2e emissions. Additionally, this weather data is utilized to analyze the relationship between each weather variable and emission patterns, providing valuable insights into the influence of weather conditions on emissions.

1. 2m Air Temperature in Celsius (v_2t): This variable measures the ambient temperature 2 meters above ground level. Temperature plays a crucial role in determining the operational efficiency of power plants. It affects the plant's cooling requirements and can influence the thermal efficiency of electricity generation, with higher temperatures potentially reducing efficiency due to increased cooling needs.
2. 2m Dew Point in Celsius (v_2d): The dew point at 2 meters indicates the moisture content in the air. Lower

dew points signify drier air, affecting coal storage and quality. Dryer conditions may reduce the moisture content of coal, potentially enhancing its combustion efficiency and affecting CO2e emissions.

3. V-wind Component on 10m in Meters per Second (v_10v): This variable measures the vertical component of wind speed at a height of 10 meters. Cyclical wind patterns can directly impact coal consumption by affecting operational cooling needs. Wind speed and direction can alter the ambient temperature around the power plant and affect the coal burned rate.
4. 10m Wind Gust in Meters per Second (v_i10fg): Wind gusts at 10 meters capture the short-term variations in wind speed. Significant variability in wind gusts, with cyclical patterns, can influence both the cooling requirements of the plant and the dispersion of emitted pollutants, including CO2 and CO2e.
5. Volumetric Soil Water at 0-7 cm in Cubic Meters (v_swvl1): This metric reflects the water content in the soil up to a depth of 7 cm, indicative of the soil's moisture level. Cyclical drops in soil moisture could be related to seasonal variations in rainfall, affecting the ambient humidity and, consequently, coal storage conditions. Higher soil moisture can increase the surrounding air's humidity, potentially impacting coal quality and emissions.

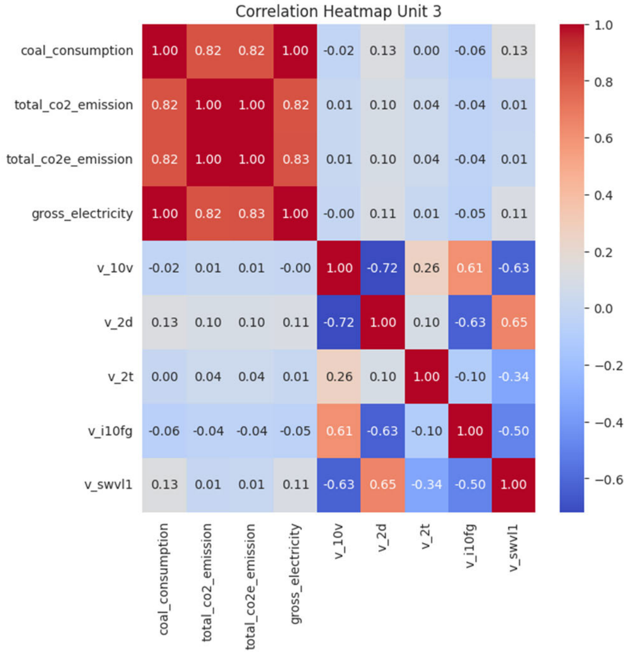


FIGURE 2. **Correlation between variables in PLTU unit 3.**

3) CORRELATION BETWEEN VARIABLES

This section explores the correlation between variables within PLTU Unit 3, Unit 7, and Unit 8. The correlation heatmap matrix for PLTU Unit 3, depicted in Figure 2, visually represents the strength and direction of the relationships among the various parameters measured. Each cell in the matrix displays the correlation coefficient between two variables, ranging from −1 to 1.

A correlation coefficient close to 1 indicates a strong positive correlation, implying that as one variable increases, the other also tends to increase. Conversely, a coefficient near −1 signifies a robust negative correlation, suggesting that as one variable increases, the other tends to decrease. A coefficient near 0 indicates a weak or no correlation between the variables.

Based on Figure 2, the correlation analysis reveals that the independent variables with the strongest correlation to the dependent variable (CO2e emissions) in PLTU Unit 3 are as follows:

1. CO2 Emissions (Correlation Value: 1): The perfect positive correlation indicates that as CO2 emissions increase or decrease, CO2e emissions also increase or decrease proportionally. This suggests a direct and substantial relationship between CO2 emissions and CO2e emissions, which is expected given that CO2 emissions contribute directly to the overall carbon footprint.
2. Gross Electricity (Correlation Value: 0.83): The strong positive correlation implies that higher levels of gross electricity generation are closely associated with increased CO2e emissions. This correlation is logical since electricity generation often involves the combustion of fossil fuels, resulting in CO2e emissions.
3. Coal Consumption (Correlation Value: 0.82): Similar to gross electricity, coal consumption demonstrates a strong positive correlation with CO2e emissions. This finding aligns with expectations, as coal is a primary energy source in power plants, and its combustion releases significant amounts of CO2e emissions.
4. Weather Variables (Highest Correlation: 2m Dew Point): Although the weather variables exhibit lower correlation values than internal variables, they still contribute to understanding emission dynamics. The highest correlation observed with the 2m Dew Point suggests that certain weather conditions, such as humidity levels, may significantly influence CO2e emissions.

Overall, the correlation analysis underscores the importance of internal factors such as CO2 emissions, gross electricity, and coal consumption in driving CO2e emissions in PLTU Unit 3. While weather variables show less significant correlations, they still provide supplementary insights into emission patterns, highlighting the multifaceted nature of emission dynamics in power plants. Next, the correlation relationships among variables in PLTU Unit 7 can be observed in Figure 3 and PLTU Unit 8 in Figure 4.

Based on Figure 3 and Figure 4, the correlation analysis conducted across all units within the PLTU reveals that the influence of variables on CO2e emissions remains relatively consistent across different units. While slight variations in the strength of correlations between specific variables may exist,

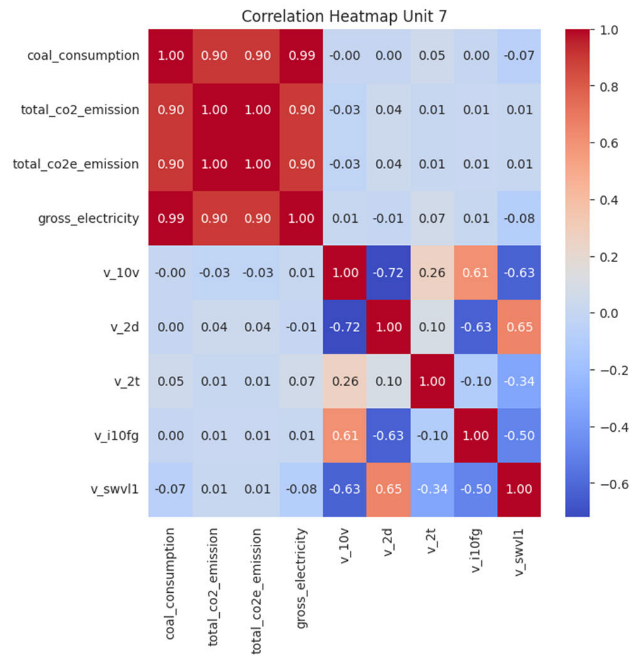


FIGURE 3. Correlation between variables in PLTU unit 7.

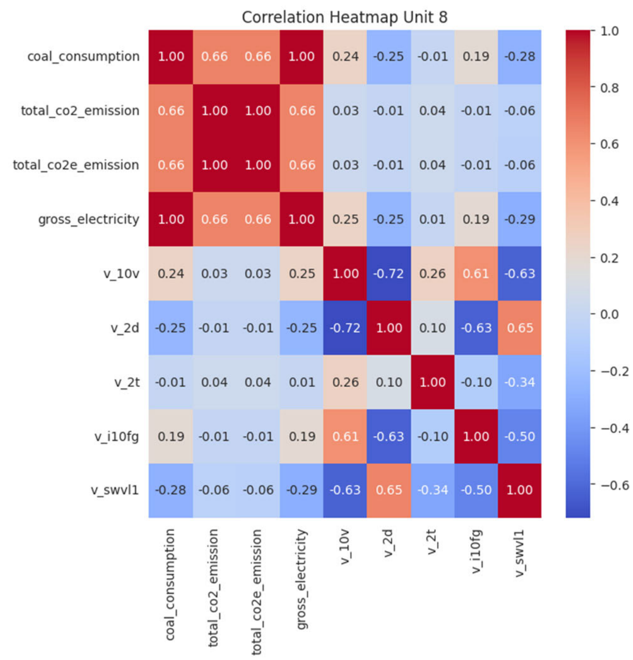


FIGURE 4. Correlation between variables in PLTU unit 8.

the overall impact and direction of influence remain broadly similar across units.

Across all units, variables such as CO2 emissions, gross electricity generation, and coal consumption consistently exhibit strong positive correlations with CO2e emissions. This suggests that these internal factors are crucial in driving carbon emissions in coal-fired power plants, regardless of the specific unit. Similarly, the influence of weather variables, although weaker than internal factors, shows a consistent pattern across units. While individual weather variables may vary in correlation strengths, they generally contribute to understanding emission dynamics within the PLTU. This uniformity in correlation patterns underscores the robustness of the analysis and indicates that the identified relationships are likely to hold across different units within the power plant. Such insights are instrumental in informing emission reduction strategies and optimizing operational efficiency across the facility.

B. DATA PREPARATION

This section will explain each stage of the Data Preparation phase. The stages in this phase start from the data preprocessing and feature engineering stages.

1) DATA PREPROCESSING

Data preprocessing is a critical stage in data preparation before it is used to train models in machine learning modeling. The details of the data preprocessing stages are as follows.

1. In time format conversion, the "time" field is converted to DateTime format using the Pandas library for better analysis and modeling.
2. Some conditions are used to delete data with invalid or irrelevant values to delete specific data. For example, data with values "Bad" and "[-11059] No Good Data For Calculation" are in the column "gross_electricity."
3. Data type change: Some data columns are changed according to the data type. For example, the column "unit_id" is converted to integer type (int), while other columns such as "coal_consumption," "total_co2_emission", and others are converted to float type, making mathematical processing easier.
4. Data grouping and aggregation: data is grouped based on the "time" column with the frequency of grouping per day (freq ="D"). Next, the aggregation function (agg_funcs) is applied to the grouped data. The grouped data is data that will be used as features such as "time," "coal_consumption," "total_co2_emission", and other data. After that, the data is reset to its index to produce aggregated data.

2) FEATURE ENGINEERING

Feature engineering is creating or changing variables in a dataset to obtain additional information or improve model performance. Feature engineering uses the Moving Average method. Moving Averages provide further information regarding trends and patterns from the data to be analyzed. The types of Moving Average used are Simple Moving Average (SMA) and Exponential Moving Average (EMA). Simple Moving Average (SMA) is a method for calculating the average of the last amount of data in a specific time range. At the same time, EMA is a method that gives greater weight to the latest data in its calculation. This makes it possible to put more emphasis on the latest trends in the data [23].

In the regression model, feature engineering is done by adding features to the "total_co2_emission", "coal_consumption," and "gross_electricity" features. SMA is calculated with windows of 5 days, 10 days, 20 days, and 30 days. Next, it is stored with a new column according to the number of windows used. Then, the EMA is calculated with filtering factors (spans) of 5 days, 10 days, 20 days, and 30 days. Next, it is stored with a new column according to the number of spans used. In the SMA method, the first index starts according to the specified window; in this case, most windows have a 30-day window, so the data on index 0-29 in window 30 will be empty, and then the data used is only data starting at index 30. The data in the regression model for PLTU Unit 3, Unit 7, and Unit 8 have been combined with previous weather data. Each unit consists of 942 rows and 10 columns or variables, but after feature engineering, it increased to 912 rows and 34 columns or variables. One sample of the correlation matrix in the regression model, namely the correlation matrix at PLTU Unit 3, can be seen in Figure 5.

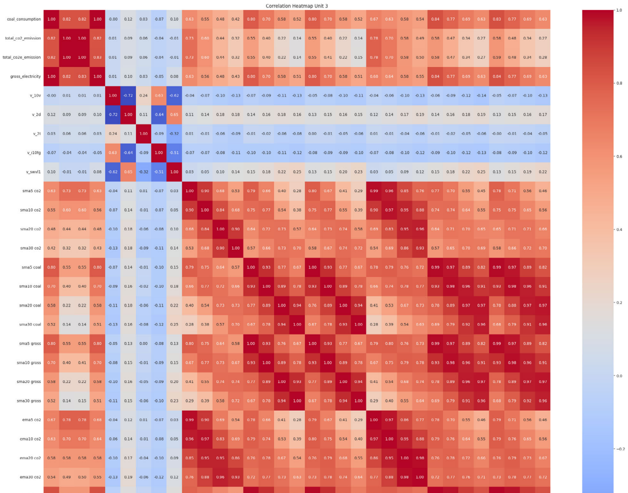


FIGURE 5. **Correlation matrix of each unit 3 regression model feature in the emission prediction model.**

In time series models or feature prediction models, the main focus of this model is to utilize historical data from each feature to predict future values, so it is essential to consider additional features that can improve prediction accuracy. Some of the techniques used are lagging, SMA, and EMA techniques. The data used is historical data from each feature or variable in the previous period. Lagging techniques carried out by the shifting process are lag 1, lag 2, lag 3, lag 4, lag 5, lag 10, lag 20, lag 30, lag 31, lag 38, lag 45, and lag 61. Furthermore, the SMA technique was carried out with windows of 2 days, 3 days, 4 days, 5 days, 10 days, 20 days, 30 days, 31 days, 38 days, 45 days, and 61 days. Then, the EMA technique is carried out with span windows of 2 days, 3 days, 4 days, 5 days, 10 days, 12 days, 14 days, 20 days, 30 days, 31 days, 38 days, 45 days, and 61 days. In this way, the model can consider the effect of emissions from the previous few days on emissions the next day, allowing for more accurate predictions and responsiveness to changes in emissions levels. One sample of the correlation matrix feature prediction, namely the feature of coal fuel consumption at PLTU Unit 3, can be seen in Figure 6.

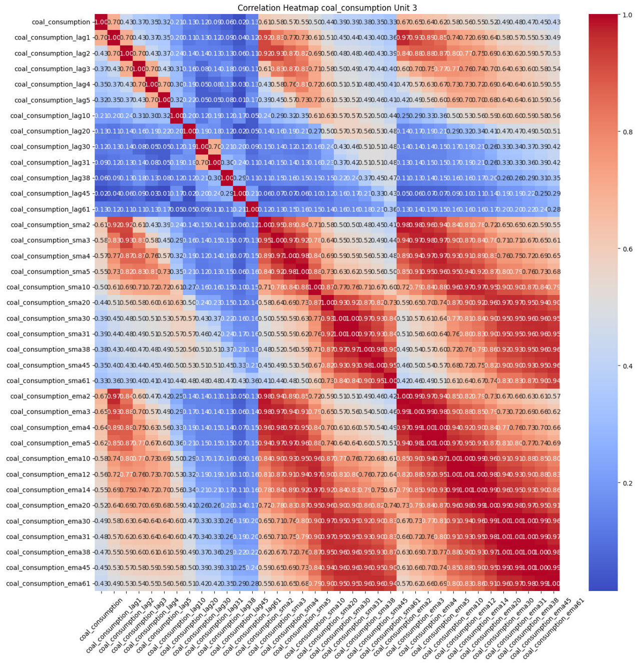


FIGURE 6. **Correlation matrix for each model feature time series unit 3 feature coal consumption.**

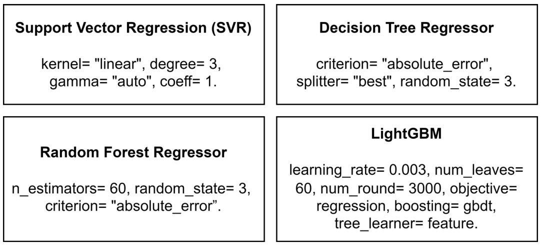
C. MODELING

Various machine learning algorithm models will be developed to assess their performance in solving specific tasks [24], [25]. Modeling techniques in time series models to predict each feature or independent variable. Furthermore, modeling techniques on carbon emission prediction models will use regression models in each generating unit. Some machine learning algorithms used in the modeling phase are as follows:

1. Linear Regression
2. Support Vector Regression (SVR)
3. Decision Tree Regression
4. Random Forest Regression
5. LightGBM

Machine learning model development involves setting up or configuring hyperparameters on each model to be developed. Hyperparameters are settings that cannot be learned by the model during the training process and must be determined before the training process begins.

Hyperparameters will affect the learning process in each model and the final performance of the model. Figure 7 represents the hyperparameters used in the time series model, while Figure 8 is the hyperparameter used in the regression model.



Support Vector Regression (SVR)

kernel= "linear", degree= 3, gamma= "auto", coeff= 1.

Decision Tree Regressor

criterion= "absolute_error", splitter= "best", random_state= 3.

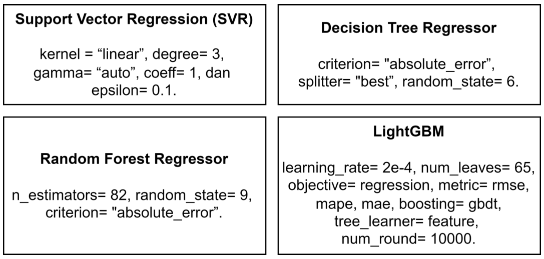
Random Forest Regressor

n_estimators= 60, random_state= 3, criterion= "absolute_error".

LightGBM

learning_rate= 0.003, num_leaves= 60, num_round= 3000, objective= regression, boosting= gbdt, tree_learner= feature.

FIGURE 7. Hyperparameters in time series models.



Support Vector Regression (SVR)

kernel = "linear", degree= 3, gamma= "auto", coeff= 1, dan epsilon= 0.1.

Decision Tree Regressor

criterion= "absolute_error", splitter= "best", random_state= 6.

Random Forest Regressor

n_estimators= 82, random_state= 9, criterion= "absolute_error".

LightGBM

learning_rate= 2e-4, num_leaves= 65, objective= regression, metric= rmse, mape, mae, boosting= gbdt, tree_learner= feature, num_round= 10000.

FIGURE 8. Hyperparameters in regression models.

D. EVALUATION

The dataset is segmented using the holdout method in the testing phase, dedicating 85% to training and 15% to testing. This distribution ensures a comprehensive learning environment for the model to identify patterns and intricacies, boosting its accuracy and generalization ability. Moreover, the allocation provides enough test data to accurately assess the model's performance on novel data, striking a balance between mitigating overfitting and evaluating effectiveness in real-world scenarios. A specific cut-off date of April 1, 2023, is set to distinctly separate the training and testing datasets, facilitating a precise evaluation of the model's predictive capabilities. This strategic split maximizes the model's exposure to diverse training examples while ensuring an impartial test set for performance validation. Choosing an 85/15 split is a calculated decision to craft a thoroughly trained model to generalize well beyond its training data, proving its value and dependability for practical applications. Testing on each model will be conducted using several metrics as follows.

1. Mean Absolute Error (MAE)
2. Root Mean Square Error (RMSE)
3. Mean Absolute Percentage Error (MAPE)

RMSE accentuates the impact of significant errors by elevating each error before averaging, while MAE offers uniform treatment of all deviations, providing a clear understanding in absolute terms. However, MAE cannot discern error magnitudes. MAE and RMSE range from 0 to infinity, contingent on the data scale [26]. In contrast, MAPE evaluates relative error without squaring mistakes, showing reduced susceptibility to outliers [27]. MAPE offers valuable insights into error percentages. Yet, it faces challenges in datasets with zero values or substantial variations, potentially yielding undefined or infinite values. Additionally, MAPE encompasses diverse assessment categories, detailed in Table 3 [28].

TABLE 3. MAPE evaluation assessment categories.

Value	Assessment Categories
<10 %	Excellent Forecasting Model Competence
10 – 20 %	Good Forecasting Model Competence
20 – 50 %	Feasible Forecasting Model Competency
>50 %	Competence of Bad Forecasting Models

E. DEPLOYMENT

The completion of model development typically marks an intermediary stage in a project. Whether the model's objective is to enhance understanding of the data or serve another purpose, end-users must structure and communicate the acquired knowledge effectively for practical utility. Frequently, this entails actively integrating the model into an organization's decision-making processes [24], [29], [30].

IV. RESULTS AND DISCUSSION

Time series and regression models have been evaluated using three metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Average Percentage Error (MAPE). The test analysis results are obtained based on the evaluation results, which will be explained in each subchapter.

A. PERFORMANCE ANALYSIS OF TIME SERIES FEATURE PREDICTION MODELING

In the time series feature prediction model, three model variations are developed based on features or independent variables that will be used as inputs to the regression model: the prediction model of coal consumption features, CO2 emissions, and gross electricity. A comparison of test results for each model on each feature can be seen in Table 4. Based on Table 4, the best model for each unit is obtained as follows.

1. Unit 3: The best model is Linear Regression with an average MAPE value of 0.14 or 14%; the value indicates the competence of the model to produce predictions in the "good" category, followed by the lowest average RMSE value of 2083.69 and the lowest average MAE value of 1604.74.
2. Unit 7: The best model is Linear Regression with an average MAPE value of 2.16 or 216%; the value shows a MAPE value that is an error because it is more than 100%; this is because the characteristic data in Unit 7 has an actual value of 0 causing the MAPE error value

because MAPE is very sensitive to the actual value of 0. However, other metrics, such as RMSE and MAE, get the lowest average value, namely the average RMSE value of 1969.54 and the average value of MAE of 1153.49.

3. Unit 8: The best model is Random Forest Regression with an average MAPE value of 0.12 or 12%; the value shows the competence of the model to produce predictions in the "good" category, followed by the lowest average RMSE value of 1271.52 and the lowest average MAE value of 1050.65.

The following are the results of the analysis based on the results of obtaining test values on the best model:

1. Linear Regression is the best model in Unit 3 and Unit 7 because the data in these units have a linear relationship between independent and dependent features. The effectiveness of Linear Regression being the best model in Unit 3 and Unit 7 shows that the data for these units have a robust linear relationship, making it easier for the model to make more accurate predictions.
2. Random Forest Regression is the best model in Unit 8 because the characteristics of the data in Unit 8 have a more complex relationship. The success of Random Forest Regression as the best model in Unit 8 shows that the data for this unit is more complicated or non-linear, so Random Forest Regression can capture this relationship better than other models.

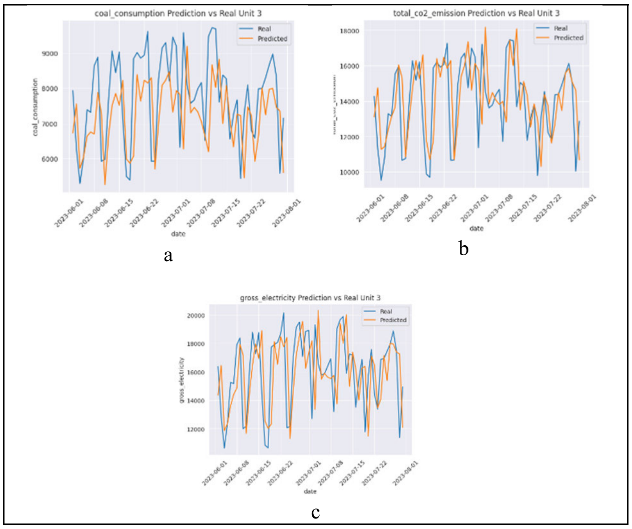


FIGURE 9. Comparison of predicted value and actual value of each feature unit 3.

A comparison of the testing dataset between the predicted value and the actual value for each feature using the Linear Regression model for PLTU Unit 3 can be seen in Figure 9. Furthermore, PLTU Unit 7, using the Linear Regression model, can be seen in Figure 10, and PLTU Unit 8, using the Random Forest Regression model, can be seen in Figure 11.

TABLE 4. Results of time series model evaluation (feature prediction).

Type	Unit	Feature	MAE	RMSE	MAPE
Linear Regression	3	Coal consumption	1081.64	1293.54	0.14
		CO2	1625.70	2221.90	0.13
		Gross electricity	2106.90	2735.63	0.14
		Average	1604.74	2083.69	0.14
Linear Regression	7	Coal consumption	727.72	1195.01	2.95
		CO2	1386.78	2487.41	1.39
		Gross electricity	1345.97	2226.21	2.14
		Average	1153.49	1969.54	2.16
Linear Regression	8	Coal consumption	678.50	829.26	0.12
		CO2	1411.04	1767.21	0.15
		Gross electricity	1279.17	1572.06	0.12
		Average	1122.90	1389.51	0.13
Support Vector Regression	3	Coal consumption	1295.18	1536.19	0.17
		CO2	1720.35	2194.66	0.13
		Gross electricity	2744.20	3453.80	0.18
		Average	1919.91	2394.88	0.16
Support Vector Regression	7	Coal consumption	762.57	1231.01	3.01
		CO2	1673.87	2515.07	1.30
		Gross electricity	1626.71	2348.23	2.19
		Average	1354.38	2031.44	2.17
Support Vector Regression	8	Coal consumption	769.16	937.97	0.14
		CO2	2848.51	3602.17	0.28
		Gross electricity	1637.22	1926.02	0.16
		Average	1751.63	2155.39	0.19
Decision Tree Regression	3	Coal consumption	1351.67	1720.87	0.18
		CO2	3073.57	3784.50	0.23
		Gross electricity	3543.74	4461.70	0.23
		Average	2656.33	3322.36	0.22
Decision Tree Regression	7	Coal consumption	2572.87	3103.55	3.38
		CO2	2063.73	3486.90	1.62
		Gross electricity	1692.02	2654.33	2.46
		Average	2109.54	3081.60	2.49
Decision Tree Regression	8	Coal consumption	954.82	1238.26	0.18
		CO2	1650.32	2437.66	0.17
		Gross electricity	2419.89	3502.14	0.21
		Average	1675.01	2392.68	0.19

TABLE 4. ***(Continued.)*** **Results of time series model evaluation (feature prediction).**

Random Forest Regression	3	Coal consumption	1019.26	1326.69	0.14
		CO2	1795.91	2337.71	0.14
		Gross electricity	2398.52	2929.23	0.16
		Average	1737.90	2197.88	0.15
Random Forest Regression	7	Coal consumption	729.02	1207.21	3.15
		CO2	1537.06	2576.73	1.43
		Gross electricity	1465.47	2324.62	2.41
		Average	1243.85	2036.18	2.33
Random Forest Regression	8	Coal consumption	623.27	791.63	0.12
		CO2	1208.96	1385.94	0.12
		Gross electricity	1319.72	1636.99	0.12
		Average	1050.65	1271.52	0.12
LightGBM	3	Coal consumption	1160.01	1480.70	0.16
		CO2	1684.52	2304.84	0.13
		Gross electricity	2397.25	3036.75	0.16
		Average	1747.26	2274.10	0.15
LightGBM	7	Coal consumption	811.75	1267.40	3.31
		CO2	1513.20	2599.08	1.43
		Gross electricity	1448.50	2306.82	2.39
		Average	1257.82	2057.77	2.38
LightGBM	8	Coal consumption	698.97	912.06	0.13
		CO2	1296.33	1576.95	0.14
		Gross electricity	1319.22	1694.12	0.12
		Average	1104.84	1394.38	0.13

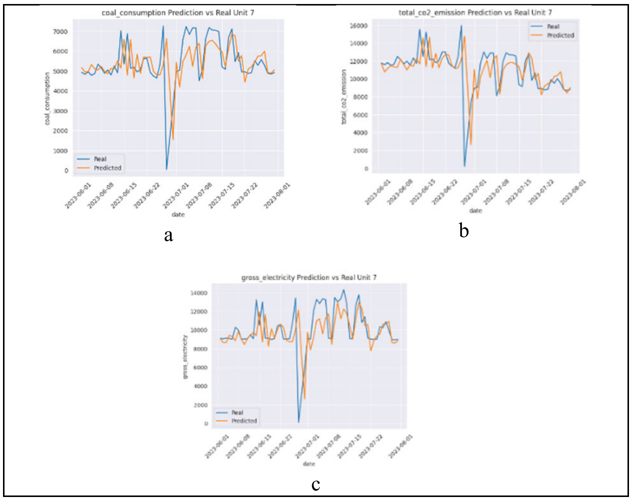


FIGURE 10. **Comparison of predicted value and actual value of each feature unit 7.**

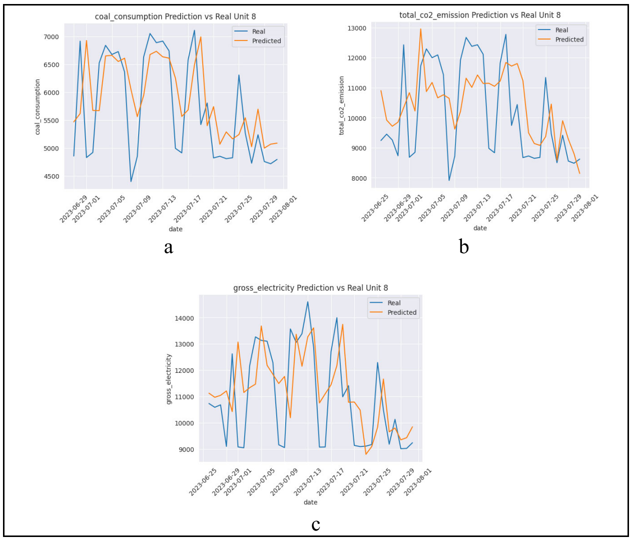


FIGURE 11. **Comparison of predicted value and actual value of each feature unit 8.**

B. PERFORMANCE ANALYSIS OF CARBON EMISSION PREDICTION MODELING

The CO2e emission prediction model uses a regression model to predict CO2e emissions based on each feature or independent variable generated from the time series feature prediction model. A comparison of test results for each model on each feature can be seen in Table 5.

Based on the test results in Table 5, the best model for all units is Linear Regression.

1. Unit 3: The best model is Linear Regression with a MAPE value of 1.7E-07 or 0.000017%; the value indicates the competence of the model to produce predictions with the ''excellent'' category, followed by the lowest RMSE value of 0.003 and the lowest MAE value of 0.002.
2. Unit 7: The best model is Linear Regression with a MAPE value of 9.2E+09; the MAPE value shows a MAPE error value because it is more than 100%; this is because the characteristic data in Unit 7 has an actual value of 0, causing the MAPE error value because MAPE is very sensitive to the actual value of 0. However, other metrics, such as RMSE and MAE, get the lowest values, namely the RMSE value of 0.002 and the MAE value of 0.001.
3. Unit 8: The best model is Linear Regression with a MAPE value of 3.2E+11; the value shows a MAPE value that is an error because it is more than 100%; this is because the characteristic data in Unit 8 has an actual value of 0 causing the MAPE error value because MAPE is very sensitive to the actual value of 0. However, other metrics, such as RMSE and MAE, get

TABLE 5. Results of regression model evaluation (CO2e emission prediction).

Type	Unit	MAE	RMSE	MAPE
Linear Regression	3	0.002	0.003	1.7E-07
	7	0.001	0.002	9.2E+09
	8	0.001	0.002	3.2E+11
Support Vector Regression	3	0.029	0.035	2.4E-06
	7	0.043	0.049	1.4E+12
	8	0.028	0.034	2.7E+13
Decision Tree Regression	3	46.162	76.168	4.2E-03
	7	39.188	73.551	4.2E-03
	8	142.308	394.711	3.8E-02
Random Forest Regression	3	21.708	33.918	1.9E-03
	7	27.073	78.008	7.5E-03
	8	81.368	163.386	6.3E+15
LightGBM	3	1510.579	1811.734	1.4E-01
	7	1226.442	1584.739	2.5E+17
	8	3939.729	4689.383	6.5E+18

the lowest values, namely the RMSE value of 0.002 and the MAE value of 0.001.

Linear Regression is the best model for all units; it is supported by a solid relationship between the features used in the model and the predicted targets, resulting in low RMSE and MAE evaluation metric values; it shows that this model is very good at making accurate predictions for all units. A comparison of the testing dataset between the predicted value of CO2e and the actual value of CO2e using the Linear Regression model in PLTU Unit 3 can be seen in Figure 12. A comparison of the testing dataset between the predicted CO2e value and actual CO2e value using the Linear Regression model in PLTU Unit 7 can be seen in Figure 13. A comparison of the testing dataset between the predicted CO2e value and actual CO2e value using the Linear Regression model in PLTU Unit 8 can be seen in Figure 14.

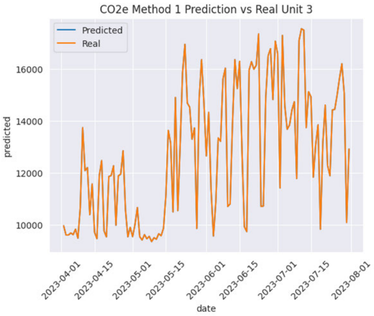


FIGURE 12. Comparison of CO2e predicted value and actual CO2e value unit 3.

C. PERFORMANCE ANALYSIS FROM THE RESULTS OF THE INTEGRATION OF THE FEATURE PREDICTION MODEL WITH THE CARBON EMISSION PREDICTION MODEL

This section analyzes the results of integrating time series models with regression to obtain the results of coal fuel consumption predictions, gross electricity predictions, and CO2e emission predictions until 2024. The results of integration will be explained in each subchapter as follows.

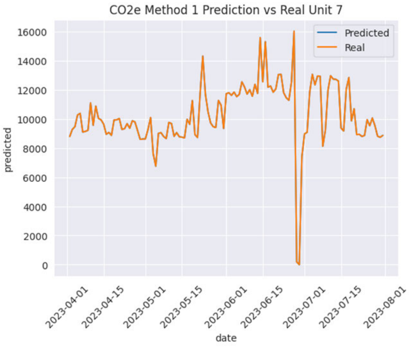


FIGURE 13. Comparison of CO2e predicted value and actual CO2e value unit 7.

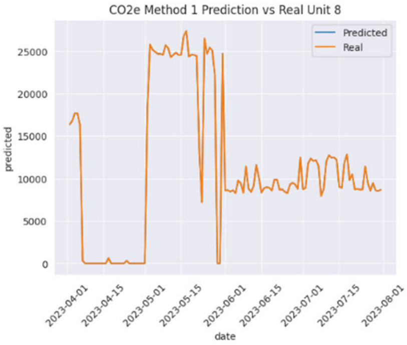


FIGURE 14. Comparison of CO2e predicted value and actual CO2e value unit 8.

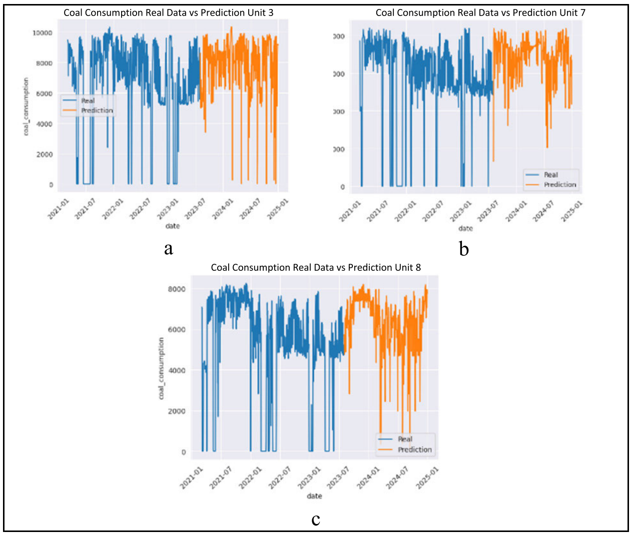


FIGURE 15. Coal fuel consumption prediction using best model.

1) ANALYSIS OF COAL CONSUMPTION PREDICTION RESULTS (TIME SERIES MODEL)

In this section, we will analyze the predicted results of coal fuel consumption based on the test results in Table 4. Figure 15 is the result of predicting coal consumption using the best model.

Based on Figure 15, the results of implementing the best model to predict coal fuel consumption in all units for 518 days or until December 30, 2024. Table 6 provides information on the prediction of coal fuel consumption using the best model from January 1, 2024, to December 30, 2024.

TABLE 6. Results of coal fuel consumption prediction in 2024.

Unit Name	Coal Fuel Consumption Prediction (tonnes)
Unit 3	2.892.427,87
Unit 7	2.518.406,35
Unit 8	2.231.703,13

Based on Table 6, the predicted results of coal fuel consumption can be used as a reference for the power plant stakeholders to estimate the use of coal fuel consumption in the future by considering the emissions produced based on the predicted results of coal fuel consumption.

2) ANALYSIS OF GROSS ELECTRICITY PREDICTION RESULTS (TIME SERIES MODEL)

This section will analyze the gross electricity prediction results based on Table 4 's test results. Figure 16 shows the results of the awful electricity prediction in Unit 3, Unit 7, and Unit 8 using the best model. Based on the results of the prediction experiment for the next 518 days or December 30, 2024. Figure 16 shows gross electricity prediction results.

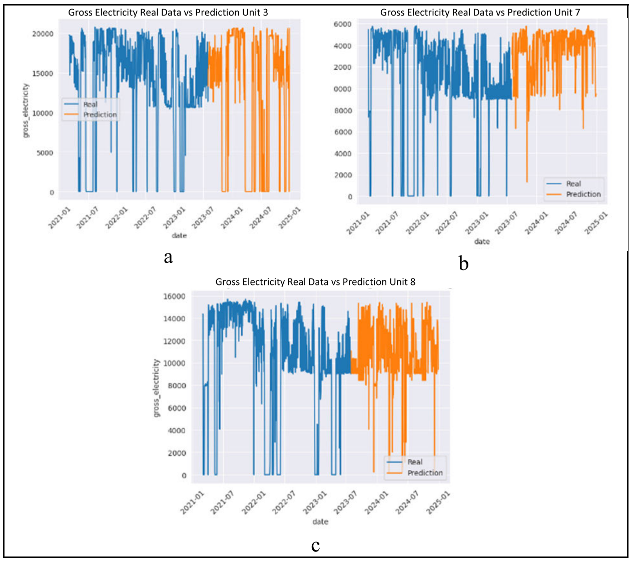


FIGURE 16. Gross electricity prediction using best model.

Figure 16 shows that it can be analyzed when the best model is implemented into predictions for as many as 518 days or until December 30, 2024. Table 7 contains gross electricity results from January 1, 2024, to December 30, 2024.

Based on Table 7, the results of gross electricity predictions can be used as a reference for power plant stakeholders to be able to estimate the results of electrical energy produced by considering the results of coal consumption predictions and emission predictions made so that the power plant can consider environmental and financial impacts in the future.

TABLE 7. Results of gross electricity prediction in 2024.

Unit Name	Gross Electricity (MW) Prediction
Unit 3	4.442.031,26
Unit 7	5.095.387,79
Unit 8	4.080.690,30

3) ANALYSIS OF CARBON EMISSION PREDICTION RESULTS (REGRESSION MODEL)

Emission prediction is obtained using each best model based on Table 5 for CO2e emission prediction in Method 1. The emission prediction results for each unit using a regression model using feature inputs based on the results of the feature prediction model using the best model. The results of the CO2e emission prediction in Method 1 can be seen in Figure 17.

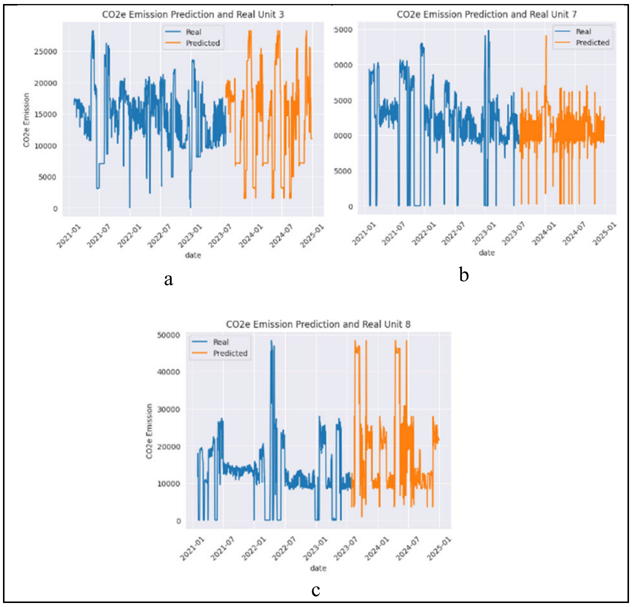


FIGURE 17. CO2e emission prediction method 1 using best model.

Based on Figure 17, Figure (a) is the result of the prediction of CO2e emissions at PLTU Unit 3 using the Linear Regression Model, Figure (b) is the result of prediction of CO2e emissions at PLTU Unit 7 using the Linear Regression Model, and Figure (c) is the result of prediction of CO2e emissions at PLTU Unit 8 using the Linear Regression Model. Implementing CO2e emission predictions using the best model results in emission predictions of 518 days or until December 30, 2024. The results of emission predictions for the 2024 period starting from January 1, 2024, to December 30, 2024, resulted in predictions of CO2e emissions in Unit 3 of 4,965,151.67 tons of CO2e, predictions of CO2e emissions in Unit 7 of 4,119,092.09 tons of CO2e, and predictions of CO2e emissions in Unit 8 of 6,397,632.42 tons of CO2e.

The findings from our predictive models, as illustrated in Figure 17, reveal valuable insights into the forecasted CO2e emissions for PLTU Units 3, 7, and 8. The high correlation observed between the independent variables and the dependent variable, as identified theoretically, underscores the significance of feature selection in regression tasks. This implies that the chosen features, including coal consumption, gross electricity, CO2 emissions, and five weather data, effectively capture the underlying dynamics influencing emissions.

Practically, these results offer actionable insights for stakeholders in coal-fired power plants. By leveraging the predictive capabilities of our models, plant operators can anticipate future emission levels with greater precision, enabling proactive decision-making in emission reduction strategies. For instance, operators can implement targeted measures to optimize plant operations and minimize environmental impact by identifying periods of high emission intensity.

Moreover, the ability to forecast emissions over an extended period, as demonstrated in our analysis until December 30, 2024, provides valuable foresight for long-term planning and regulatory compliance. This allows power plant stakeholders to align their strategies with emission reduction targets and carbon trading regulations, promoting sustainability and environmental responsibility.

V. CONCLUSION

In this study, we have developed a hybrid predictive model that integrates time series and regression techniques to forecast CO2e emissions in coal-fired power plants. Time series models forecast individual variables used as features for CO2e emissions prediction, subsequently employed as inputs for regression models. By combining internal and external data sources, including coal consumption, gross electricity, CO2 emissions, and five weather data, our model provides a comprehensive understanding of CO2e emissions dynamics.

Evaluation metrics such as MAE, RMSE, and MAPE demonstrate the model's effectiveness, with Linear Regression performing optimally in time series modeling for PLTU Units 3 and 7 and Random Forest Regression proving superior for PLTU Unit 8. Subsequently, the regression model testing results for all units produced the best model, Linear Regression. This underscores the importance of considering data characteristics and selecting appropriate models for accurate predictions.

Furthermore, our study highlights the significance of incorporating time series data into regression models. Unlike traditional regression models, which rely solely on independent variables, including time series data enables the prediction of future independent variable values, which is crucial for forecasting CO2e emissions. This methodology empowers power plants to anticipate carbon emissions and make informed decisions regarding carbon trading. Additionally, coal consumption and gross electricity predictions provide valuable insights into fuel requirements and environmental impacts, facilitating effective emission reduction strategies aligned with regulatory requirements and operational goals. However, bridging the gap between research findings and real-world implementation is crucial to ensure practical applicability.

Future work should focus on integrating our research results with practical applications in Indonesian power plants. This would enable the optimization of carbon trading decisions and emission reduction efforts, thereby contributing to the advancement of sustainable practices in the power generation industry. Future research should explore additional machine learning models beyond those examined in this study and incorporate regular data training updates to capture evolving trends effectively. Alternative calculation methods, such as Continuous Emission Monitoring Systems (CEMS), could enhance the holistic emission forecasting and management approach.

In conclusion, our hybrid predictive model framework enhances emission prediction accuracy and facilitates informed decision-making regarding carbon trading and emission reduction strategies in coal-fired power plants. Our study advances sustainable practices in the power generation industry by leveraging advanced data mining techniques, such as time series and regression modeling.

ACKNOWLEDGMENT

The authors thank Institut Teknologi Bandung and Telkom University for their valuable support in facilitating this research.

REFERENCES

[1] S. F. Pileggi and S. A. Lamia, "Climate change TimeLine: An ontology to tell the story so far," *IEEE Access*, vol. 8, pp. 65294–65312, 2020, doi: 10.1109/ACCESS.2020.2985112.

[2] E. Conticini, B. Frediani, and D. Caro, "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?" *Environ. Pollut.*, vol. 261, Jun. 2020, Art. no. 114465, doi: 10.1016/j.envpol.2020.114465.

[3] M. Emami Javanmard and S. F. Ghaderi, "A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data," *Sustain. Cities Soc.*, vol. 82, Jul. 2022, Art. no. 103886, doi: 10.1016/j.scs.2022.103886.

[4] World Nuclear Association. (2022). *Carbon Dioxide Emissions From Electricity*. [Online]. Available: https://world-nuclear.org/information-library/energy-and-the-environment/carbon-dioxide-emissions-from-electricity.aspx

[5] H. Ritchie, M. Roser, and P. Rosado. (2020). *CO_2 and Greenhouse Gas Emissions*. [Online]. Available: https://ourworldindata.org/co2-and-greenhouse-gas-emissions

[6] *Inventory of GHG Emissions in the Energy Sector*, Ministry Energy Mineral Resour. Republic Indonesia, Indonesia, 2020.

[7] Ministry of Energy and Mineral Resources of the Republic of Indonesia. (2023). *Press Release Number 022. Pers/04/SJI/2023 on the Ministry of Energy and Mineral Resources Issues Reference Values R Carbon Economic Sub-Sector of Electricity*. [Online]. Available: https://www.esdm.go.id/id/media-center/arsip-berita/kementerian-esdm-terbitkan-aturan-acuan-nilai-ekonomi-karbon-subsektor-listrik

[8] Energy Sector, Sub-Sector of Electricity, *Guidelines for Calculation and Reporting of Greenhouse Gas Inventory*, Ministry Energy Mineral Resour. Republic Indonesia, Indonesia, 2018.

[9] Government of Indonesia, *Regulation Number 16 of 2022 on the Procedures for Organizing the Economic Value of Carbon in the Sub-Sector of Power Generation*, Ministry Energy Mineral Resour. Republic Indonesia, Indonesia, 2022.

[10] E. Shabani, B. Hayati, E. Pishbahar, M. A. Ghorbani, and M. Ghahremanzadeh, "A novel approach to predict CO_2 emission in the agriculture sector of Iran based on inclusive multiple model," *J. Cleaner Prod.*, vol. 279, Jan. 2021, Art. no. 123708, doi: 10.1016/j.jclepro.2020.123708.

[11] V. Aryai and M. Goldsworthy, "Day ahead carbon emission forecasting of the regional national electricity market using machine learning methods," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106314, doi: 10.1016/j.engappai.2023.106314.

[12] Y. K. Dwivedi, L. Hughes, A. K. Kar, A. M. Baabdullah, P. Grover, R. Abbas, D. Andreini, I. Abumoghli, Y. Barlette, D. Bunker, and L. C. Kruse, "Climate change and COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action," *Int. J. Inf. Manag.*, vol. 63, Apr. 2022, Art. no. 102456, doi: 10.1016/j.ijinfomgt.2021.102456.

[13] N. A. Herawati, K. Surendro, and E. Hikmawati, "Development of an intelligent carbon emission monitoring methodology to support carbon trading with a design thinking approach: A case study in PT paiton energy," in *Proc. 10th Int. Conf. ICT Smart Soc. (ICISS)*, Bandung, Indonesia, Sep. 2023, pp. 1–6, doi: 10.1109/iciss59129.2023.10291249.

[14] M. Hans, E. Hikmawati, and K. Surendro, "Predictive analytics model for optimizing carbon footprint from students' learning activities in computer science-related majors," *IEEE Access*, vol. 11, pp. 114976–114991, 2023, doi: 10.1109/ACCESS.2023.3324725.

[15] P. Kadam and S. Vijayumar, "Prediction model: CO_2 emission using machine learning," in *Proc. 3rd Int. Conf. Converg. Technol. (I2CT)*, Pune, India, Apr. 2018, pp. 1–3, doi: 10.1109/I2CT.2018.8529498.

[16] C. Zhu, P. Shi, Z. Li, M. Li, H. Zhang, and T. Ding, "Carbon emission prediction of thermal power plants based on machine learning techniques," in *Proc. 5th Int. Conf. Energy, Electr. Power Eng. (CEEPE)*, Chongqing, China, Apr. 2022, pp. 1142–1146, doi: 10.1109/CEEPE55110.2022.9783417.

[17] H. Tan, "Carbon emission prediction with macroeconomic variables and machine learning," in *Proc. 3rd Int. Conf. Clean Green Energy Eng. (CGEE)*, Istanbul, Turkey, Aug. 2022, pp. 52–56, doi: 10.1109/CGEE55282.2022.9976625.

[18] Z. Huang, C. Huang, and Z. Wen, "Comparison of carbon emission forecasting in Guangdong province based on multiple machine learning models," in *Proc. IEEE 5th Int. Conf. Knowl. Innov. Invention (ICKII)*, Hualien, Taiwan, Jul. 2022, pp. 90–93, doi: 10.1109/ICKII55100.2022.9983576.

[19] N. Zheng, W. Chen, S. Chen, J. Chen, H. Chen, and X. Lin, "Research on carbon emission prediction method of power systems considering unit coal consumption," in *Proc. IEEE 6th Conf. Energy Internet Energy System Integr. (EI)*, Chengdu, China, Nov. 2022, pp. 2666–2670, doi: 10.1109/EI256261.2022.10116780.

[20] *Carbon Trading Trial at Coal-Fired Power Plant*, Ministry Energy Mineral Resour. Republic Indonesia, Indonesia, 2021.

[21] *Ministerial Decision Number 141.K/TL.05.DJL.4/2023 Dated*, Ministry Energy Mineral Resour. Republic Indonesia, Indonesia, Jan. 2023.

[22] Meteum AI. (2023). *Historical Weather Data: Location Probolinggo Jawa Timur*. [Online]. Available: https://b2b.meteum.ai/b2b/api/history_archive

[23] A. Kumila, B. Sholihah, E. Evizia, N. Safitri, and S. Fitri, "Perbandingan metode moving average dan metode Naïve dalam peramalan data kemiskinan," *JTAM J. Teori dan Aplikasi Matematika*, vol. 3, no. 1, pp. 65–73, Apr. 1, 2019, doi: 10.31764/jtam.v3i1.764.

[24] J. A. M. Fadhlurrahman, N. A. Herawati, H. R. W. Aulya, I. Puspasari, and N. P. Utama, "Sentiment analysis of game reviews on STEAM using BERT, BiLSTM, and CRF," in *Proc. Int. Conf. Elect. Eng. Inform. (ICEEI)*, Bandung, Indonesia, 2023, pp. 1–6, doi: 10.1109/ICEEI59426.2023.10346219.

[25] A. K. Oktavius, S. R. Manalu, Y. Indrianti, and J. V. Moniaga, "Artificial intelligence in entrepreneurial mindfulness using CRISP-DM method," in *Proc. IEEE 7th Int. Conf. Inf. Technol. Digit. Appl. (ICITDA)*, Nov. 2022, pp. 1–7, doi: 10.1109/ICITDA55840.2022.9971384.

[26] Amazon Web Services. (2023). *Metrics and Validation*. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-metrics-validation.html

[27] A. Stephen. (2023). *RMSE vs MAPE, Which is the Best Regression Metric*. [Online]. Available: https://stephenallwright.com/rmse-vs-mape/

[28] I. Nabillah and I. Ranggadara, "Mean absolute percentage error untuk evaluasi hasil prediksi komoditas laut," *J. Inf. Syst.*, vol. 5, no. 2, pp. 250–255, Nov. 2020, doi: 10.33633/joins.v5i2.3900.

[29] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Proc. Comput. Sci.*, vol. 181, pp. 526–534, Jan. 2021, doi: 10.1016/j.procs.2021.01.199.

[30] W. Y. Ayele, "Adapting CRISP-DM for idea mining: A data mining process for generating ideas using a textual dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020, doi: 10.14569/ijacsa.2020.0110603.

NENG AYU HERAWATI received the B.Ed. degree in education from Universitas Negeri Jakarta, in 2021, and the master's degree in informatics from Institut Teknologi Bandung (ITB), in 2024. She is currently a Lecturer Assistant or an Academic Tutor with ITB. Her research interests include information systems, data science, and machine learning.

ASYRAF ATTHARIQ PUTRA GARY received the B.Ed. degree in education from Universitas Negeri Jakarta, in 2021, and the master's degree in electrical engineering from Institut Teknologi Bandung (ITB), in 2023. He is currently an Independent Researcher and preparing to pursue the Ph.D. degree. His research interests include data science, analysis, defense studies, and virtual reality development.

ERNA HIKMAWATI received the B.S. degree in computer science from the National University of Pasim, in 2013, the master's degree in information systems from STMIK LIKMI, in 2017, and the Ph.D. degree in electrical engineering and informatics from Institut Teknologi Bandung (ITB), in 2023. She is currently a Lecturer of application software engineering with the School of Applied Science, Telkom University. Her research interests include data science, machine learning, and information systems.

KRIDANTO SURENDRO (Member, IEEE) received the B.Eng. and master's degrees in industrial engineering from Institut Teknologi Bandung (ITB), Indonesia, in 1987 and 1991, respectively, and the Ph.D. degree in computer science from Keio University, Japan, in 1999. He is currently a computer science professor with ITB. His research interests include data science, soft computing, and IT governance.