Journal of Theoretical and Applied Information Technology
31st December 2024. Vol.102. No. 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# PREDICTION OF POWER OUTPUT ON EXTERNAL COMBUSTION ENGINE USING REGRESSION MODELS

**RIZQI FITRI NARYANTO[1], MERA KARTIKA DELIMAYANTI[2], RIZKY ADI[3], ABDURRAHMAN[4], PUTRI KHOIRIN NASHIROH[5], IMAM SUKOCO[6], FIQRI FADILLAH FAHMI[7], DAMAI YUDHA AKBAR EFFENDI[8], AFRILZA DAFFA NARYAPRAMONO[9]**

[1,6,8]Mechanical Engineering Department, Engineering Faculty, Universitas Negeri Semarang, Semarang, Indonesia
[2,3]Department of Computer and Informatics Engineering, Politeknik Negeri Jakarta, Depok, Indonesia
[5]Informatic Engineering Education Department, Engineering Faculty, Universitas Negeri Semarang, Semarang, Indonesia
[4,7]Automotive Engineering Education Department, Engineering Faculty, Universitas Negeri Semarang, Semarang, Indonesia
[9]Informatic Engineering, Mathematics and Natural Sciences Faculty, Universitas Negeri Semarang, Semarang, Indonesia

E-mail: [1]rizqi_fitri@mail.unnes.ac.id

**ABSTRACT**

This study explores the application of machine learning regression models to predict power output in External Combustion Engine on Combined Cycle Power Plants (CCPPs) using a comprehensive dataset of 9,568 hourly observations from 2006 to 2011. Key ambient variables include temperature, pressure, humidity, and vacuum. To prevent overfitting, a 5x2 fold cross-validation strategy is employed, generating 10 unique training and testing sets. Several models are assessed, including Random Forest, XGB Regressor, Extra Trees, Hist Gradient Boosting, and LGBM Regressor. XGB Regressor demonstrates superior performance with a Mean Absolute Error (MAE) of 2.41 and Root Mean Squared Error (RMSE) of 3.37, making it the most accurate model. Additionally, the performance of ensemble models further highlights their reliability in predicting power output. The study emphasizes the importance of advanced machine learning techniques in optimizing power predictions, balancing computational efficiency, accuracy, and interpretability for large-scale industrial applications. Boosting Regressor provides a more equitable compromise between computational efficiency and performance, rendering it well-suited for implementations on a large scale. Furthermore, despite its marginally diminished accuracy, the Random Forest Regressor offers significant insights via the feature importance analysis, thereby augmenting interpretability. This study underscores the significance of sophisticated machine learning models in enhancing the precision and effectiveness of power output forecasts in CCPPs. It stresses balancing interpretability, computational cost, and accuracy in real-world applications.

**Keywords:** *Cross-Validation, External Combustion Engine, Machine Learning Models, Power Output, Performance Metrics*

## 1. INTRODUCTION

The application of machine learning across various fields has redefined the traditional approach to those areas by enhancing the process through data-based operation. The use of machine learning algorithms can provide extraordinary accuracy when analyzing huge amounts of operational data and power output estimations[1], [2]. This could result in much better energy efficiencies and greater overall reliability. This topic is well illustrated by the efforts to increase power output from engines based on external combustion. The estimation of performance and the management of such machines should also be accurate enough to reduce the amount of energy that can be saved. Besides, such machines can be a vital component in many industry procedures [3]. The figure of the External Combustion Engine (ECE) was shown in figure 1 that indicates that an external combustion engine system is where fuel is

burned outside the actual engine to provide heat. This heat will later be used in a boiler, turning water into steam. The steam produced will drive a turbine and a generator, producing electricity. Steaming out of the turbine, the steam goes into the condenser, which cools and condenses into liquid water. This water is then pumped back to the boiler, repeating the cycle. Again, since this is external combustion, a wide range of fuels can be used, and the process usually has lower emissions than internal combustion engines, in which fuel is directly burned inside the engine cylinders.

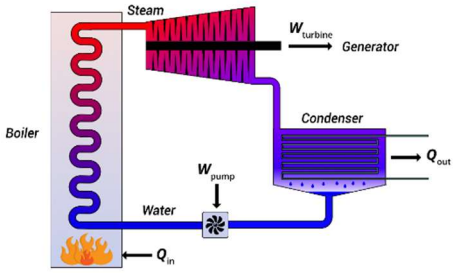


*Figure 1 The External Combustion Engine*

Management of complex non-linear relationships in machine performance data can be accomplished with the help of machine learning, which offers a robust framework. Even in the presence of noise and unexpected data points, machine learning algorithms can acquire information from previous data, identify patterns, and deliver accurate projections. Approaches considered conventional frequently require appropriate consideration of the loads of elements and interactions that affect the functioning of the machine [4]. As a result of this research, it has been demonstrated that the application of machine learning techniques, such as neural networks and support vector machines, can significantly enhance the precision of power output estimations compared to the conventional methods. These features are essential when it comes to improving the operating parameters of an external combustion engine because they guarantee that the engine will work within its ideal performance range while simultaneously reducing the amount of fuel consumed and emissions produced [5], [6].

Several studies have explored regression techniques to model power output in various energy systems, including internal combustion engines and renewable energy systems[7]. Regression models in machine learning are employed to forecast continuous values by analyzing the interconnections among variables. Regression analysis is a statistical technique used to create a mathematical model representing the relationship between independent and dependent variables. It aids in comprehending the relationship between the dependent variable and the independent variables by observing how the value of the dependent variable changes with different values of the independent variables[8]. For example, [9] applied multiple linear regression (MLR) to predict performance in diesel engines, while [10] demonstrated the superiority of polynomial regression in capturing complex interactions among variables in renewable energy setups. However, research specifically targeting the application of regression models to external combustion engines remains limited, leaving a critical gap in the understanding of how these techniques can be optimized for External Combustion Engine performance prediction. Regression models are extensively utilized in diverse applications, including predicting continuous outcomes such as house prices, stock prices, or sales. They are also employed to estimate the performance of upcoming retail sales or marketing initiatives and predict client or user trends [11]. The selection of a regression model is contingent upon the characteristics of the data and the specific problem being tackled. Some common regression models are simple linear regression, multiple linear regression, logistic regression, and polynomial regression [12].

Regression models are most effective when the target variable is continuous, and there is a distinct relationship between the predictor and target variables. They are extensively utilized in banking, marketing, and other domains where accurate forecasting and prediction are critical [11], [13], [14]. Organizations can enhance their performance by employing regression models to generate forecasts and make better decisions by comprehending the correlations between variables. The regression analysis method has been shown to have excellent prediction accuracy and reliability in energy output prediction models. As a result, it is a significant tool for optimizing the performance of external combustion engines. It is a very successful approach of machine learning to use regression analysis, particularly when employing tools for regression analysis such as XGBoost, LightGBM, and Extremely Randomized Trees. The primary purpose for which these algorithms were developed is to process big data sets that contain many dimensions and intricate interactions between variables [2], [12], [15]. As a result of its excellent scalability and performance, XGBoost is well-known for its capacity to manage big data sets with minimal errors [16]. The LightGBM algorithm, on

the other hand, is highly effective and can provide predictions that are both quick and accurate while consuming less memory [17]. Decision Trees, well-known for their ease of use and reliable performance, also perform exceptionally well in regression situations because they reduce variation through randomization [18].

It is possible to quantify the average size of the error in a particular group of forecasts using the Mean Absolute Error (MAE) metric, regardless of the direction of the error. One can use this as a straightforward and straightforward indicator of how accurate the predictions are. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are two examples of metrics typically utilized to evaluate machine learning models' efficiency. On the other hand, the root mean square error (RMSE) is more dependent on more significant errors and more sensitive to outliers. The phrase provided is the square root of the average of the squared discrepancies between the expected value and the value that occurred. Evaluating the predictive performance of machine learning models in energy production projections requires both processes to be completed. This ensures that the algorithms offer estimates that are both dependable and accurate at the same time. By utilizing these evaluation criteria, researchers can ensure that their models are accurate and resistant to fluctuations in the data it contains [19], [20].

In addition to playing a significant part in accomplishing the overarching objective of reducing energy consumption, machine learning enhances the capability of predicting the requirements for maintenance and developing efficient operational plans. By utilizing machine learning techniques, businesses can achieve sustainability in their operations, reduce their operational expenses, and contribute to protecting the environment. In this circumstance, the significance of utilizing machine learning cannot be stressed sufficiently. Improving the efficiency of energy systems is necessary considering the growing demand for energy worldwide [2], [15], [21]. The results of previous studies demonstrate that machine learning has the potential to minimize energy use by removing operational inefficiencies and encouraging environmentally friendly choices from individuals [3], [5]. The purpose of this study is to evaluate the capability of machine learning to forecast the amount of power generated by external combustion engines [22]. This will allow for the creation of energy management solutions that are both more intelligent and efficient.

## 2. METHOD

### 2.1 Dataset

The Combined Cycle Power Plant (CCPP) dataset, which is extensively employed in power generation prediction tasks and resulted from the external combustion engine, is utilized in this study. The CCPP dataset comprises operational data obtained from a combined cycle power plant, where Gas Turbines (GT), Steam Turbines (ST), and Heat Recovery Steam Generators (HRSG) measurements are integrated. Combined cycle power stations are specifically engineered to optimize operational effectiveness through the synchronized utilization of gas and steam turbines [5], [23]. The dataset presents a distinctive architecture consisting of five randomized iterations of the data, each encompassing 9568 data points gathered on an hourly basis for a period of six years, from 2006 to 2011. The structure of this comprehensive dataset is tailored to accommodate the 5x2 fold cross-validation method, which is a reliable approach for assessing the performance of models in machine learning endeavors. By employing this methodology, which consists of dividing the dataset into five discrete subsets and performing two iterations of cross-validation, the model's precision and applicability are comprehensively evaluated [23].

Every data point in the dataset is composed of four characteristics, which are the hourly mean values of the surrounding conditions. Vacuum (V), Ambient Temperature (T), Atmospheric Pressure (AP), and Relative Humidity (RH) are these characteristics. The ambient temperature (T) indicates the temperature of the external air, which has the potential to substantially impact the efficiency of the power facility. The pressure exerted by the weight of the atmosphere, known as Atmospheric Pressure (AP), has the potential to influence the combustion process within gas turbines. The RH value serves as an indicator of the relative humidity, which influences the refrigeration mechanisms operating within the facility. The term vacuum (V) denotes the critical pressure differential that exists between the steam condenser and the atmosphere, which has a direct bearing on the efficacy of the steam turbine [24]. The target variable in this dataset is the continuous net hourly electrical power output (PE) of the facility; therefore, this is a regression task. It is essential to forecast the net hourly electrical power output to improve the efficiency and operation of the power

facility. The capacity to precisely predict power output in accordance with ambient conditions enhances resource management at the power facility and contributes to more consistent and effective power generation [24].

For researchers and practitioners in the fields of machine learning and power systems, the CCPP dataset is indispensable. It furnishes an extensive array of characteristics and a demanding prediction endeavor that can be employed to evaluate a multitude of machine learning algorithms and methodologies. The dataset facilitates the creation of more precise and dependable power output prediction models, a critical aspect in ensuring the effective functioning of combined cycle power plants—by providing an extensive and detailed arsenal of operational data. In summary, the CCPP dataset serves as an indispensable asset in the progression of power generation prediction. The exhaustive data and well-organized format of this system establish a solid foundation for assessing and enhancing machine learning models, which ultimately contributes to the development of power generation systems that are more dependable and efficient [12], [24].

### 2.2 Training Pipeline

The research technique utilizes a conventional deep-learning training process to construct and assess regression models to predict power output. The training pipeline follows a sequential process to get the most optimal model. The training methodology utilized in this study is depicted in Figure 2.

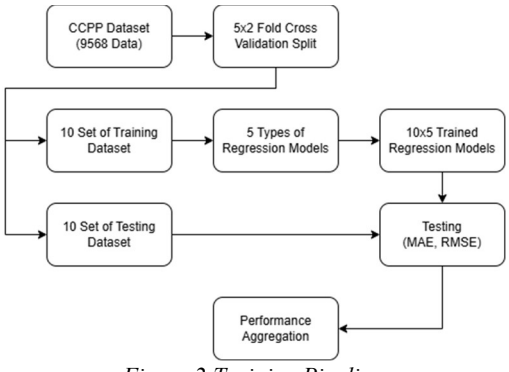


*Figure 2 Training Pipeline*

The research technique utilizes a conventional deep-learning training process to construct and assess regression models to predict power output. The training pipeline follows a sequential process to get the most optimal model. The training methodology utilized in this study is depicted in Figure 2. Figure 2 shows our pipeline for model training in this research. We start by using CCPP dataset as our dataset[5], [22]. After that, we applied 5x2 fold cross validation split, which resulting in 10 distinct set of training dataset and 10 distinct set of testing dataset. This cross-validation technique was employed to make sure a robust evaluation of model generalization capability. These training set then used to train five different regression models, which are XGBRegressor, ExtraTreesRegressor, HistGradientBoostingRegressor, LGBMRegressor, and RandomForestRegressor. These models were chosen to represent a diverse range of ensemble learning, trees, and gradient boosting algorithm techniques, enabling a comprehensive comparison of their predictive capabilities. This training step resulting in 50 trained models, which corresponds to 10 set of training dataset for each 5 regression models. Each trained model then tested using the testing dataset, focusing on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as the evaluation metrices, following the standard metric of regression tasks [20]. This process is repeated for all 10 train-test splits of 5 regression models to obtain a comprehensive model performance evaluation. Lastly, the final performance of each model aggregated by averaging the MAE and RMSE values across all 10 evaluation instances. This provides a robust and unbiased estimate of the model's predictive performance.

During the second stage, known as model training, five distinct regression models are trained using each training set. The models employed in this study consist of XGBRegressor [16], [17], [25], [26], ExtraTreesRegressor [27], [28], HistGradientBoostingRegressor [27], LGBMRegressor [17] and RandomForestRegressor [29], [30]. These models encompass various ensemble learning and gradient-boosting strategies, facilitating a thorough evaluation of their prediction abilities. XGBRegressor and LGBMRegressor are renowned for their exceptional efficiency and superior performance in gradient boosting. ExtraTreesRegressor and RandomForestRegressor are resilient ensemble algorithms that mitigate overfitting by averaging the predictions of numerous decision trees. HistGradientBoostingRegressor is a high-speed and precise implementation of gradient boosting that efficiently handles big datasets.

Once the models have been trained, the subsequent step involves evaluating the models. The

performance of each training model is assessed on the appropriate test set using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), commonly used metrics for regression tasks. Mean Absolute Error (MAE) quantifies the average size of mistakes in each set of forecasts, offering a clear indication of the accuracy of the predictions. Root Mean Square Error (RMSE) quantifies the square root of the mean of the squared discrepancies between predicted and actual values. It assigns a greater weight to more significant errors, highlighting precise predictions' significance. This exhaustive evaluation approach is iterated for all ten train-test splits to guarantee a meticulous assessment of each model's performance across diverse data partitions [20], [21], [31].

In the final stage, performance aggregation, the overall performance of each model is determined by calculating the average of the MAE and RMSE values over all ten evaluation instances. This yields a strong and impartial evaluation of the model's ability to make accurate predictions. By calculating the average performance metrics across numerous splits, we ensure that the findings are not excessively impacted by any specific train-test split, resulting in a more dependable comparison of the models.This precise and systematic approach to developing and evaluating models enables the discovery of the most efficient regression model for predicting power production in combined cycle power plants. By employing numerous models and conducting thorough cross-validation, the study assures that its conclusions are strong and applicable to various situations. This provides vital insights into the ability of different machine learning techniques to make accurate predictions in this specific context [2], [15].

Furthermore, it is vital to comprehend the importance of each model employed in this investigation and these scientific procedures. The XGBRegressor is renowned for its implementation of the extreme gradient boosting technique and its ability to handle a combination of continuous and categorical data effectively [32]. LGBMRegressor, a modified version of gradient boosting, is designed to prioritize speed and performance, making it well-suited for handling extensive datasets [17]. The ExtraTreesRegressor algorithm constructs numerous decision trees and combines their predictions, providing resilience against overfitting [27], [28]. RandomForestRegressor is an ensemble method that enhances predictive accuracy and mitigates overfitting by averaging the predictions of numerous decision trees [27], [29]. Lastly, the HistGradientBoostingRegressor is an accelerated variant of gradient boosting that utilizes histogram-based techniques to improve processing efficiency [27].

The methodology emphasizes the significance of a well-organized training procedure and thorough evaluation in creating highly efficient machine learning models for intricate tasks like power output prediction. This research enhances the field of predictive modelling in power production by utilizing a wide range of models and comprehensive evaluation approaches. It can potentially enhance operational efficiency and decision-making in combined cycle power plants [2], [33]. Overall, the methodology's methodical approach guarantees that the final models are accurate and generalizable, able to produce dependable forecasts for power production applications in the real world.

### 2.3 Cross-Validation Strategy

To avoid overfitting and guarantee a dependable assessment of the model's ability to generalize, this research employs a rigorous 5x2 fold cross-validation approach. Cross-validation is essential for assessing machine learning models, significantly when the dataset is constrained or susceptible to bias. Within this particular framework, the Combined Cycle Power Plant (CCPP) dataset poses a distinctive obstacle because of its operational complexities and the imperative for precise power generation forecasting [24]. To tackle these obstacles, the research utilizes the innate shuffles in the CCPP dataset and implements a 5x2 fold cross-validation methodology. The procedure entails partitioning the dataset into five randomized iterations, followed by a two-fold cross-validation procedure for each iteration. The outcome is the generation of ten distinct training and testing sets, which facilitate the assessment of the model on various subsets of data.

The model's robustness and dependability are significantly improved by this diversity, which is essential for evaluating its performance under various conditions and scenarios. By implementing cross-validation across numerous shuffles of the dataset, the research guarantees a comprehensive assessment of the model's performance across a wide range of data distributions [34], [35]. This methodology reduces the likelihood of overfitting, in which the model memorizes the training data instead of making accurate predictions on unobserved data. Furthermore, it furnishes a more pragmatic assessment of the model's efficacy in practical scenarios, wherein data distributions might fluctuate over time or among distinct operational environments. Following the completion of cross-

validation on every iteration of the dataset, the research compiles the outcomes to depict the ultimate performance of the model. By implementing this aggregating procedure, any discrepancies in performance among distinct data subsets are mitigated, resulting in a more consistent and dependable estimation of the predictive prowess of the model. The study obtains a comprehensive evaluation of the model's overall performance, including accuracy and error rates, by calculating the mean of the performance metrics [34], [36], [37]. This assessment is vital to make informed decisions regarding the model's feasibility for practical implementation.

### 2.4 Regression Models

Five distinct regression models has been considered to predict the power output target data, and their models are:

1. XGBRegressor: a model which uses the gradient boosting framework, where the weak learners are decision trees, and is capable to capture complex non- linear relationships between features and target variable with high accuracy [16], [25], [26], [32].

2. ExtraTreesRegressor: this model uses an ensemble of randomized decision trees that introduces additional randomness into feature permutation during splits and helps in generalizing and preventing overfitting [27], [28].

3. HistGradientBoostingRegressor: a model that uses a histogram-based approach for gradient boosting which provides computational efficiency while maintaining the level of predictivity to be competitive, generally used with large-scale data [27].

4. LGBMRegressor: stands for Light Gradient Boosting Machine Regressor is a model that uses the light gradient boosting framework, optimized for fast computation speed and efficient memory usage[17].

5. RandomForestRegressor: a model that involves ensemble learning, averaging multiple decision tree predictions due to its resistance to outliers and effectiveness in capturing intricate feature relationships [27], [29].

### 2.5 Evaluations Metrics

Model evaluation is an essential component of machine learning, particularly in regression tasks such as estimating power output in a combined cycle power plant (CCPP). This study utilizes two main metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), to evaluate the effectiveness of regression models. Mean Absolute Error (MAE) is a measure that quantifies the average absolute difference between projected and actual data. It offers a concise comprehension of the mean forecast inaccuracy. The Mean Absolute Error (MAE) formula, represented by Eq. (1), calculates the sum of the absolute differences between each anticipated value ($\hat{y}_i$) and its corresponding actual value ($y_i$). This sum is then divided by the total number of data points (n). The Mean Absolute Error (MAE) provides a straightforward and transparent measure of model performance by calculating prediction mistakes without considering their direction[38].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

Where n is the total of data. $y_i$ is the actual value, and $\hat{y}_i$ is predicted value.

Conversely, RMSE calculates the square root of the mean squared difference between expected and actual values. RMSE, unlike MAE, applies a greater penalty to more significant errors, thereby making it more responsive to outliers. RMSE is commonly employed in regression tasks to assess prediction accuracy holistically. The Root Mean Square Error (RMSE) formula, as expressed in Eq.(2), computes the square root of the average of the squared discrepancies between each predicted value ($\hat{y}_i$) and its corresponding actual value ($y_i$), divided by the total number of data points (n). The Root Mean Square Error (RMSE) calculates the average magnitude of mistakes by taking the square root of the average squared differences. This metric provides a measure of the overall forecast accuracy of the model. The user's text is a single period [38], [39].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Where n is the total of data. $y_i$ is the actual value, and $\hat{y}_i$ is predicted value.

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used as quantitative metrics to evaluate the accuracy of regression models in predicting power output. They provide essential information on the precision and dependability of the model's forecasts, assisting stakeholders in making informed decisions.

Researchers can determine the optimal model configuration for a specific job by analyzing and comparing alternative models' MAE and RMSE values or tuning parameters. Moreover, these metrics aid in choosing and implementing models by offering explicit and comprehensible measures of predictive effectiveness. MAE and RMSE are crucial evaluation measures in regression tasks, such as predicting power output in CCPP. Their computation yields valuable data regarding the predictive abilities of regression models, allowing researchers to make well-informed judgments and enhance model performance.

## 3. RESULTS AND DISCUSSION

The results of our study demonstrate the effectiveness of various regression models in predicting power output using the CCPP dataset from external combustion engine [21], [38]. Table 1 presents a comparative analysis of the model performance based on the average MAE and RMSE values obtained through the 5x2 fold cross-validation process.

*Table 1 Dataset Label Distribution*

| Model | MAE | RMSE |
|---|---|---|
| Bagging REP Tree [23] | 2.82 | 3.79 |
| XGBRegressor | 2.41 | 3.37 |
| ExtraTreesRegressor | 2.53 | 3.54 |
| HistGradientBoostingRegressor | 2.56 | 3.49 |
| LGBMRegressor | 2.56 | 3.49 |
| RandomForestRegressor | 2.56 | 3.54 |

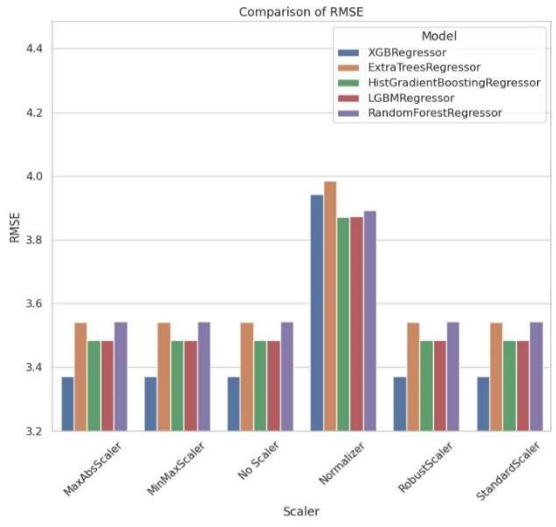


*Figure 3 The result of Scaler for every model*

As shown in Table 1 above, the study explores the relative effectiveness of different regression models in forecasting power output using environmental factors. The XGBRegressor stands out as the best performance of all these models, with the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values. The results demonstrate that XGBRegressor well captures the complex non-linear connections between ambient variables and power output, outperforming other models such as ExtraTreesRegressor, HistGradientBoostingRegressor, LGBMRegressor, and RandomForestRegressor [21].

Figure 3 had shown the effect of scaler parameter for every regressor model A scaler is a configuration instrument in a regression model employed to standardize the features (input variables) before model training. Scaling aims to standardize all features to a uniform range or distribution. This can enhance the performance of machine learning models, particularly those sensitive to input data size. In regression models, a scaler ensures uniformity among components by standardizing or normalizing features. In regression models, employing a scaler is essential, as features with disparate scales can adversely affect the model's performance and training stability. Specific algorithms, such as linear regression, exhibit sensitivity to the magnitudes of the input features. Variables having larger values can exert a more significant influence on the model outcomes. Scaling ensures that all features contribute uniformly. This is particularly crucial for enhancing algorithms such as gradient descent, as it accelerates their performance. Scaling enhances numerical stability, hence reducing the likelihood of errors resulting from significant value disparities. It facilitates a more precise and uniform comprehension of model coefficients, resulting in enhanced analysis and insight into the significance of elements inside the regression model[40], [41].

A comparison with a previous model further demonstrates the better performance of every model in the experiment, the Bagging REP Tree. This highlights the strength and efficiency of the chosen models, especially in the case of XGBRegressor. Its accurate power output prediction is due to its ability to capture non-linearities and interactions between features efficiently. Nevertheless, it is crucial to acknowledge that ensemble learning techniques like RandomForestRegressor, ExtraTreesRegressor, and HistGradientBoostingRegressor also exhibit

resilience, albeit with significantly lower performance than XGBRegressor.

Factors beyond the predictions' accuracy become essential when putting something into practice. Although XGBRegressor may provide more outstanding performance, other considerations, such as computational efficiency and interpretability, are also important. For instance, the HistGradientBoostingRegressor algorithm is often used because it can compromise performance and computational cost. This makes it well-suited for large-scale applications that have limited computational resources. However, RandomForestRegressor, while slightly less accurate than XGBRegressor, offers interpretability by examining the importance of features. The ability to interpret the results can be highly beneficial when it is essential to comprehend the fundamental aspects that affect forecasts to make informed decisions.

The exceptional efficacy of XGBRegressor, specifically in capturing intricate non-linear connections and feature interactions, highlights the need to utilize sophisticated methodologies in regression modelling, particularly in domains defined by intricate and ever-changing relationships between variables. Nevertheless, it is essential to recognize that selecting the most suitable model for actual implementation relies on several aspects, such as computing efficiency, interpretability, and the specific needs of the use case [26], [32].

This paper identifies many problems in implementing machine learning models for predicting power production in Combined Cycle Power Plants (CCPPs). Principal concerns encompass the intricacy of reconciling precision with computing efficiency, susceptibility to dataset attributes, and scalability under resource-limited settings[6], [24]. Furthermore, the trade-off between interpretability and performance remains a critical challenge for industrial applications. Future research should improve computationally intensive models for real-time applications, integrate various datasets to enhance generalizability and investigate hybrid and explainable AI models. Integrating machine learning with Computer Vision, IoT and edge computing frameworks[42] has substantial potential for enhancing predictive skills in energy systems.

## 4. CONCLUSION AND FUTURE WORKS

This study illustrates the effectiveness of advanced machine learning regression techniques in forecasting the power output of Combined Cycle Power Plants (CCPPs) that employ an external combustion engine. XGBRegressor demonstrated superior performance to other models, recording the lowest Mean Absolute Error (MAE) of 2.41 and a Root Mean Squared Error (RMSE) of 3.37. The capacity to capture intricate non-linear relationships between ambient variables and power output provides a significant advantage compared to alternative methods. While XGBRegressor demonstrates high accuracy, its computational cost presents challenges for real-time applications, indicating a necessity for further optimization in practical contexts.

The research confirms the effectiveness of ensemble learning methods, specifically ExtraTreesRegressor, HistGradientBoostingRegressor, LGBMRegressor, and RandomForestRegressor, which yielded consistent predictions. These methods exhibited statistically significant enhancements compared to prior studies' Bagging REP Tree model, highlighting their superior performance. XGBRegressor demonstrated superior accuracy; however, HistGradientBoostingRegressor proved to be a more computationally efficient option, rendering it appropriate for large-scale applications. The trade-off between accuracy and computational efficiency illustrates the essential balance required for practical implementation. The study has limitations that suggest directions for future research. The dataset, while comprehensive, was obtained from a single power plant, which raises concerns regarding the generalizability of the findings to various operational contexts. The emphasis on ambient conditions as predictive variables neglects other potential factors, including fuel type, maintenance schedules, and equipment age, that may affect power output. Addressing these gaps may improve the accuracy and applicability of machine learning models in power generation.

Future research should explore the creation of hybrid models that integrate the interpretability of RandomForestRegressor with the accuracy of gradient-boosting models such as XGBRegressor. Furthermore, optimizing computationally intensive models via pruning or model compression may facilitate their application in real-time systems. Incorporating diverse power plants with different environmental and operational attributes into datasets is essential for validating the scalability and reliability of these models. This study highlights the

efficacy of advanced machine learning models, specifically ensemble and gradient-boosting techniques, in enhancing power output prediction in combined cycle power plants (CCPPs). Compared to more straightforward regression methods, these models provide enhanced accuracy, operational efficiency, and improved managerial decision-making. Addressing current limitations and exploring open research issues may facilitate advancements in machine learning that enhance energy systems' sustainability, resilience, and efficiency, thereby contributing to the overarching objective of global energy sustainability.

## ACKNOWLEDGMENT

We would like to express our highest gratitude to Universitas Negeri Semarang for supporting this research through Fundamental Research Schema 2024.

## REFERENCES

[1] R. F. Naryanto, M. K. Delimayanti, A. Naryaningsih, B. Warsuta, R. Adi, and B. A. Setiawan, "Diesel Engine Fault Detection using Deep Learning Based on LSTM," in *2023 7th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, Dec. 2023, pp. 37–42. doi: 10.1109/ELTICOM61905.2023.10443110.

[2] R. F. Naryanto and M. K. Delimayanti, "Machine learning approach for prediction model on biomass characteristic analysis," presented at the INTERNATIONAL CONFERENCE ON MECHANICAL ENGINEERING FOR EMERGING TECHNOLOGIES (ICOMEET 2021), Padang, Indonesia, 2023, p. 040007. doi: 10.1063/5.0115617.

[3] S. Alketbi, A. B. Nassif, M. A. Eddin, I. Shahin, and A. Elnagar, "Predicting the power of a combined cycle power plant using machine learning methods," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, Sharjah, United Arab Emirates: IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/CCCI49893.2020.9256742.

[4] R. Chodkiewicz, B. Donevski, J. Krzton, and J. Porochnicki, "RECUPERATED GAS TURBINE INTEGRATED WITH THE PFBC COMBINED CYCLE POWER PLANT," *Strojniški vestnik - Journal of Mechanical Engineering*, vol. 47, no. 8, p. 6, 2001.

[5] H. Kaya, P. Tufekci, and Gürgen, Firket S., "Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine," in *International Conference on Emerging Trends in Computer and Electronics Engineering (ICETCEE'2012)*, Mar. 2012.

[6] R. F. Naryanto, M. K. Delimayanti, A. Naryaningsih, R. Adi, and B. A. Setiawan, "FAULT DETECTION IN DIESEL ENGINES USING ARTIFICIAL NEURAL NETWORKS AND CONVOLUTIONAL NEURAL NETWORKS," . *Vol.*, no. 2.

[7] R. F. Naryanto *et al.*, "An aerodynamic analysis of energy saving car based on computational fluid dynamic using solidworks," presented at the ANNUAL SYMPOSIUM ON APPLIED AND INNOVATION TECHNOLOGICAL ENVIRONMENT 2023 (ASAITE2023): Smart Technology based on Revolution Industry 4.0 and Society 5.0, Jakarta, Indonesia, 2024, p. 050008. doi: 10.1063/5.0230191.

[8] S. Rong and Z. Bao-wen, "The research of regression model in machine learning field," *MATEC Web Conf.*, vol. 176, p. 01033, 2018, doi: 10.1051/matecconf/201817601033.

[9] A. B. K. Didavi, R. G. Agbokpanzo, and M. Agbomahena, "Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, Dec. 2021, pp. 1–5. doi: 10.1109/BioSMART54244.2021.9677566.

[10] B. Cai *et al.*, "Fault detection and diagnostic method of diesel engine by combining rule-based algorithm and BNs/BPNNs," *Journal of Manufacturing Systems*, vol. 57, pp. 148–157, Oct. 2020, doi: 10.1016/j.jmsy.2020.09.001.

[11] V. Weerakkody, U. Sivarajah, K. Mahroof, T. Maruyama, and S. Lu, "Influencing subjective well-being for business and sustainable development using big data and predictive regression analysis," *Journal of Business Research*, vol. 131, pp. 520–538, Jul. 2021, doi: 10.1016/j.jbusres.2020.07.038.

[12] A. Saxena, R. Chauhan, D. Chauhan, D. S. Sharma, and D. D. S. and V. Narayan, "Comparative Analysis Of AI Regression And Classification Models For Predicting House

Damages İn Nepal: Proposed Architectures And Techniques," *Journal of Pharmaceutical Negative Results*, pp. 6203–6215, Dec. 2022, doi: 10.47750/pnr.2022.13.S10.767.

[13] R. Jiang, Y. Xin, Z. Chen, and Y. Zhang, "A medical big data access control model based on fuzzy trust prediction and regression analysis," *Applied Soft Computing*, vol. 117, p. 108423, Mar. 2022, doi: 10.1016/j.asoc.2022.108423.

[14] C. Maheswari, E. B. Priyanka, S. Thangavel, S. V. R. Vignesh, and C. Poongodi, "Multiple regression analysis for the prediction of extraction efficiency in mining industry with industrial IoT," *Prod. Eng. Res. Devel.*, vol. 14, no. 4, pp. 457–471, Oct. 2020, doi: 10.1007/s11740-020-00970-z.

[15] E. A. Elfaki and A. H. Ahmed, "Prediction of Electrical Output Power of Combined Cycle Power Plant Using Regression ANN Model," *JPEE*, vol. 06, no. 12, pp. 17–38, 2018, doi: 10.4236/jpee.2018.612002.

[16] J. Dong, Y. Chen, B. Yao, X. Zhang, and N. Zeng, "A neural network boosting regression model based on XGBoost," *Applied Soft Computing*, vol. 125, p. 109067, Aug. 2022, doi: 10.1016/j.asoc.2022.109067.

[17] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 28, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[18] F. Elmaz, Ö. Yücel, and A. Y. Mutlu, "Predictive modeling of biomass gasification with machine learning-based regression methods," *Energy*, vol. 191, p. 116541, Jan. 2020, doi: 10.1016/j.energy.2019.116541.

[19] K. D. Ismael and S. Irina, "Face recognition using Viola-Jones depending on Python," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 3, pp. 1513–1521, 2020, doi: 10.11591/ijeecs.v20.i3.pp1513-1521.

[20] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Front. Earth Sci.*, vol. 9, p. 761990, Oct. 2021, doi: 10.3389/feart.2021.761990.

[21] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, *Investigation of performance metrics in regression analysis and machine learning-based prediction models*. European Community on Computational Methods in Applied Sciences, 2022. doi: 10.23967/eccomas.2022.155.

[22] Institut Teknologi Sepuluh Nopember, J. Fadil, S. Soedibyo, Institut Teknologi Sepuluh Nopember, M. Ashari, and Institut Teknologi Sepuluh Nopember, "Novel of Vertical Axis Wind Turbine with Variable Swept Area Using Fuzzy Logic Controller," *IJIES*, vol. 13, no. 3, pp. 256–267, Jun. 2020, doi: 10.22266/ijies2020.0630.24.

[23] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power & Energy Systems*, vol. 60, pp. 126–140, Sep. 2014, doi: 10.1016/j.ijepes.2014.02.027.

[24] P. Tüfekci, and H. Kaya, "Combined Cycle Power Plant Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant#

[25] X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, "Predicting Missing Values in Medical Data Via XGBoost Regression," *J Healthc Inform Res*, vol. 4, no. 4, pp. 383–394, Dec. 2020, doi: 10.1007/s41666-020-00077-1.

[26] R. Wang, L. Wang, J. Zhang, M. He, and J. Xu, "XGBoost Machine Learning Algorism Performed Better Than Regression Models in Predicting Mortality of Moderate-to-Severe Traumatic Brain Injury," *World Neurosurgery*, vol. 163, pp. e617–e622, Jul. 2022, doi: 10.1016/j.wneu.2022.04.044.

[27] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, Art. no. 1, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[29] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[30] J. C. Pedraza, O. A. Romero, and H. E. Espitia, "Prediction of Atmospheric Pollution Using Neural Networks Model," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 6574–6581,

2020, doi: 10.11591/IJECE.V10I6.PP6574-6581.

[31] A. Botchkarev, "A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045–076, Jan. 2019.

[32] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[33] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," *IEEE International Conference on Big Data (Big Data)*, 2019.

[34] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," in *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 83–87. doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[35] K. S. Raju, M. R. Murty, and M. V. Rao, "Support Vector Machine with K-fold Cross Validation Model for Software Fault Prediction," p. 15.

[36] H. Hamdani, H. R. Hatta, N. Puspitasari, A. Septiarini, and Henderi, "Dengue classification method using support vector machines and cross-validation techniques," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1119–1129, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp1119-1129.

[37] Y. Nie, L. De Santis, M. Carratu, M. O'Nils, P. Sommella, and J. Lundgren, "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Bari, Italy: IEEE, Jun. 2020, pp. 1–6. doi: 10.1109/MeMeA49120.2020.9137222.

[38] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Information Sciences*, vol. 585, pp. 609–629, Mar. 2022, doi: 10.1016/j.ins.2021.11.036.

[39] M. Ćalasan, S. H. E. Abdel Aleem, and A. F. Zobaa, "On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function," *Energy Conversion and Management*, vol. 210, p. 112716, Apr. 2020, doi: 10.1016/j.enconman.2020.112716.

[40] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[41] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[42] W. Prastiwinarti, M. K. Delimayanti, H. Kurniawan, Y. P. Pratama, H. Wendho, and R. Adi, "Efficient packaging defect detection: leveraging pre-trained vision models through transfer learning," *IJEECS*, vol. 34, no. 3, p. 2096, Jun. 2024, doi: 10.11591/ijeecs.v34.i3.pp2096-2106.