Optimization of random forest for active power prediction based on three-phase voltage and current parameters

Muhammad Dani Solihin[1], Erita Astrid[2], Muchsin Harahap[3], Mhd Ikhsan Rifki[4], M. Khalil Gibran[5], Amir Saleh[6]

{mdnsolihin@unimed.ac.id[1], rifki.mhdikhsan@uinsu.ac.id[4], m.khalil1100000202@uinsu.ac.id[5], amirsaleh@polmed.ac.id[6]}

Department of Electrical Engineering Education, Faculty of Engineering, Universitas Negeri Medan, North Sumatra, Indonesia[1], Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, North Sumatra, Indonesia[4,5], Department of Computer Engineering and Informatics, Politeknik Negeri Medan, North Sumatra, Indonesia[6]

Abstract. This study aims to apply random forest optimization in predicting total active power in a three-phase power system based on voltage and current parameters for each phase. The data used consists of measurement data collected using power quality meters, with a total of 2,343. The target output of the research focuses on the total active power value with a variable range of 0 - 4,915.05 kW, with a data division ratio of 80% for training and 20% for testing. The model scenario was configured using the Randomized Search CV optimization method, which produced regression evaluation metrics with an MAE of 5,911.61 watts, an RMSE of 74,308.10, and a determination coefficient R² of 0.8927. The model visualization is displayed in a scatter diagram, error histogram, and comparison graph of actual and predicted active power, showing that the model provides a stable error distribution and has a good level of capability. The results of the study indicate that the random forest model with hyperparameter optimization can be used to model the non-linear patterns and characteristics of data between voltage and total active power current parameters.

Keywords: Random Forest Regression, Active Power Prediction, Voltage and Current Features , Three-Phase Electrical System, Model Optimization

1 Introduction

Energy consumption has increased significantly in line with industrial development. This has led to the emergence of more optimal, accurate, and reliable power management systems. In the large-scale industrial sector, the majority implements a 3-phase electrical power system concept due to the operational requirements of equipment and machinery, which are the mainstay of production activities. Therefore, three-phase electrical system parameters such as voltage and current play a central role in determining active power. Active power is used as a key indicator that describes the level of electrical energy that is actually converted to load. Therefore, accuracy in determining active power predictions is crucial in large-scale

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361154

industries. This is because errors in active power predictions can lead to increased operational loads, energy waste, and potentially reduced electrical system reliability.

In several relevant studies, the application of methods in making predictions is expressed in the form of mathematical and statistical models that are implemented in the process of predicting electricity consumption. Classical methods such as linear regression and time series analysis have been used, but they have limitations in capturing nonlinear data patterns and complex interactions between variables [1]. Therefore, with developments in the field of artificial intelligence and machine learning, these have begun to be implemented to provide a more accurate representation of the results for model prediction applications. Several studies indicate that the implementation of methods with a machine learning approach, such as Support Vector Machine (SVM), Neural Network, and decision tree, can provide prediction results with better accuracy compared to classical and conventional methods. One of the machine learning algorithms often used in prediction models is Random Forest. Fundamentally, Random Forest is a development of the ensemble learning method. The working mechanism of Random Forest is described as using a majority of decision trees with the bagging principle and random feature selection, thereby minimizing the potential for overfitting and potentially improving model generalization.

Research conducted by [2] describes a model built by implementing Random Forest in air-based photovoltaic-thermal (PV/T) system modeling, which provides superior evaluation results in predicting electrical and thermal efficiency with R² values of 99.99% and 81.2%, respectively. while the MAE visualization results provided results of 0.0034 and 2.64 in predicting electrical and thermal efficiency. Meanwhile, [3] ely overcome inherent variability and non-linear characteristics of wind patterns. The Random Forest approach used provides advantages that are visualized in the model evaluation results, where Random Forest has the best performance with an RMSE value of 55.11 and an R² value of 0.9882, outperforming the Neural Network, XGBoost, and linear regression models.

In improvising to achieve efficiency in machine learning, optimization is necessary to find solution options that lead to optimal answers. Optimization algorithms can be implemented to meet these objectives so that the resulting accuracy has a better value [4]. Optimization implementation in random forests can be done using several methods, including Randomized Search CV [5], because random forests are known to excel in multidimensional data classification, but their performance is highly dependent on hyperparameter optimization. Thus, the optimization performed provides a 14% improvement in accuracy over the previous results. Optimization options in random forests can also be done by implementing the Bayes algorithm. In study [6], the Bayes algorithm was used to optimize hyperparameters in creating a vulnerability assessment model for landslides using random forests. Meanwhile, study [7] utilized optimization through GridSearchCV to classify text in hoax news. The evaluation results detected an increase in accuracy from 96% to 97% in the random forest model.

The implementation of random forests has been widely used in electricity load prediction, or even energy consumption prediction, in the industrial sector. However, specifically, its optimization in predicting active power in three-phase electrical systems is still limited. Most studies concentrate on predicting energy consumption over a long period of time. Therefore, it is necessary to optimize active power in three-phase electrical systems, which are characterized by high system complexity and variability. This study focuses on the application of random forest optimization in predicting active power in three-phase electrical systems. Optimization is carried out by searching for the best parameters that can improve the random forest prediction results. The performance of the random forest will be evaluated using Mean Absolute Error (MAE), Root Mean Square, and the coefficient of determination (R^2). The contribution of this research includes a prediction model for active power in three-phase electrical systems, accompanied by a comparative analysis between the random forest and its optimization. In addition, it provides recommendations for approaches to monitoring active power in industrial-scale electrical systems.

2 Research Methodology

The application of random forest and optimization involves several processes and stages. Figure 1 illustrates the process diagram for creating a prediction model using random forest.

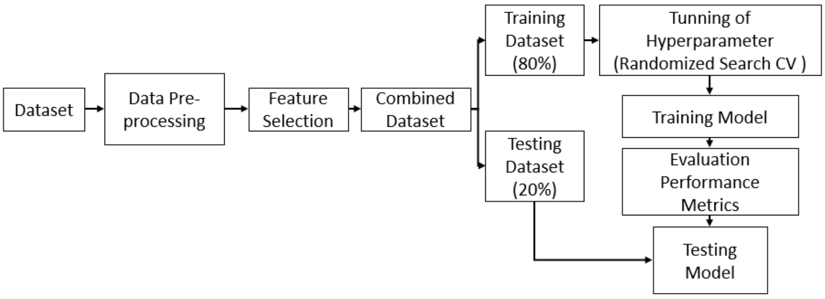


Fig. 1. Research process diagram

The dataset used in this study consists of 2,343 data points, with a number of variables, including vrms, current, phase angle, active power, reactive power, apparent power, and power factor. The dataset was obtained from Kaggle, after which data preprocessing was carried out by scanning for empty data and data anomalies. The preprocessed data has the same number of data points as the original dataset. The process continued with data splitting, with 80% of the data used for training (a total of 1,874) and 20% for testing (a total of 469).

The process continued with hyperparameter tuning on the random forest model using the Randomized Search CV method. The hyperparameter configuration is described in Table 1.

Table 1. Hyperparameter table for randomized search cv

No	Parameter Configuration	Randomized Search CV
1	*n_estimators*	100-1000
2	*max_depth*	10,20
3	*min_samples_split*	2-20
4	*min_samples_leaf*	1-10
5	*max_features*	Number of features

After that, the data was trained and evaluated using a regression evaluation matrix covering three aspects, including:

1. R-Square (R2)

   R² is used to evaluate the proportion of variance in the independent variable in a regression model, thereby determining the suitability of the model used. R2 can be calculated using the following equation [2] :

   $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_\imath})^2}{\sum_{i=1}^{n}(y_i - \widehat{y_\imath})^2} \tag{1}$$

   The R2 result indicator describes the value interval from 0 to 1. If the R2 calculation result is close to 1, then the model is considered to have satisfactory capability in explaining data variation.

2. Mean Ansolute Error (MAE)

   MAE is applied to represent the absolute average between the prediction value level and the observation value level of the dependent variable. MAE computation can be performed using the following equation [8] :

   $$MAE \ = \frac{1}{n} - \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \tag{2}$$

   The MAE computation result indicator represents the statement that the smaller the MAE computation result, the better the model applied.

3. Root Mean Square Error (RMSE)

   RMSE is used in evaluating regression models with sensitivity to outliers and is greatly affected by large errors in predictions. RMSE computation can be implemented using the following equation [9]

   $$RMSE \ = \sqrt{\frac{1}{n} - \sum_{i=1}^{n} |y_i - \hat{y}_i|^2} \tag{3}$$

   The RMSE computation result indicator is represented by a small value, so if the value obtained has a small scale, the model is considered to have satisfactory performance.

Each variable y_i in a series of evaluation matrix equations represents the actual value in the i-th data. Meanwhile, $\widehat{y_\imath}$ represents the predicted value in the i-th data. Meanwhile, the variable n represents the number of observations.

3 Results and Discussion

The dataset used in the study includes several parameters required in active power computation, including voltage, current, power factor, and active power level. An illustration of the first 200 rows of the dataset can be seen in Figure 2.

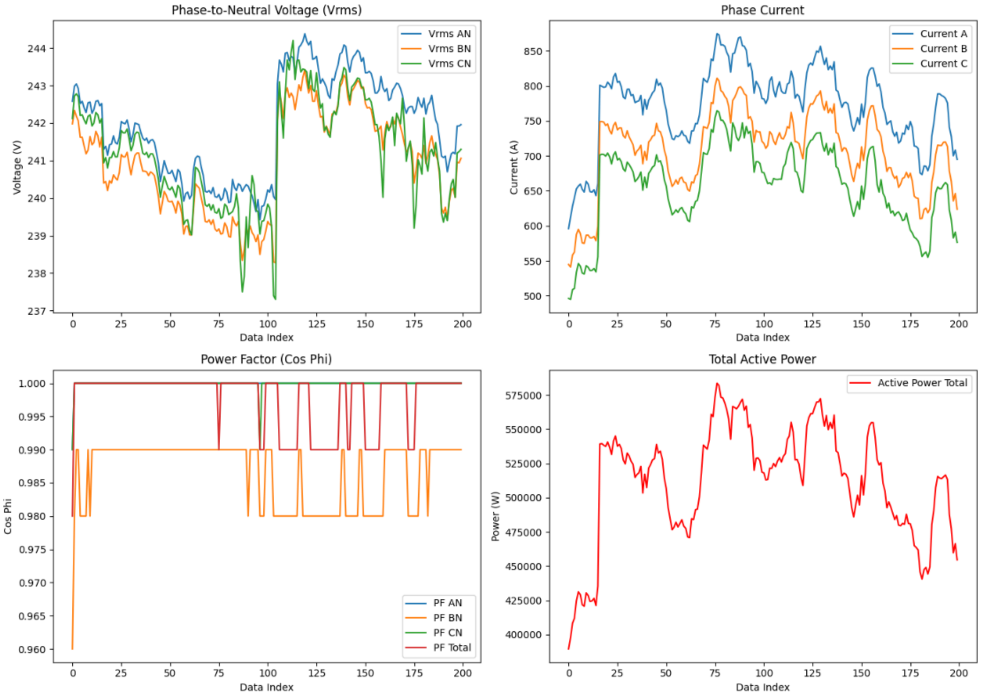
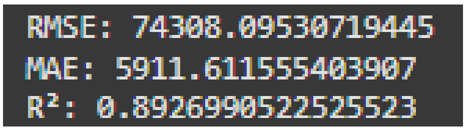


Fig. 2. Research dataset sample

The process continues with the implementation of feature selection, which is used to determine the dominant features that greatly influence computation and active power prediction processes in electrical systems. Some of the selected and used features include voltage, current, power factor, and actual active power. This procedure is linearly related to the active power computation pattern, so features that are not included in these requirements will be removed. The process continues by combining all the specified features.

The evaluation results were obtained by comparing several levels of active power values between the random forest optimization predictions and the actual active power found in the dataset. Figure 3 illustrates the prediction results from using random forest optimization with the Randomized Search CV method.



```
RMSE: 74308.09530719445
MAE: 5911.611555403907
R²: 0.8926990522525523
```

Fig. 3. Random forest optimization prediction model evaluation matrix computation results

Based on the results of the regression evaluation matrix computation in Figure 3, the description of the evaluation results analysis can be explained as follows:

1. The RMSE implemented in measuring the average active power prediction error has a value level of 73.308,10, which describes the average error of the prediction results made by the model.
2. MAE is used to determine the average absolute error between the model's computational prediction and the actual active power value. MAE produces a value of 5.911,61, which describes the average deviation of the prediction results.
3. R^2 indicates the proportion of active power variation explained by the model. The R^2 calculation yields a value of 0,8927, indicating the model's ability to understand the relationship between features and targets.

However, considering that the active power range obtained from the dataset ranges from 0 to 4.915.050 W, the model built to make predictions has an error percentage of 1,51%, which describes a low average error level. Prediction errors occur at several points of value variants that are considered to deviate significantly from the actual values of the available dataset. Meanwhile, the MAE percentage has an error percentage of 0,12%. This value is still in the small percentage category. Finally, the R^2 value, which is close to 1, indicates that the model is almost 90% capable of detecting active power variations. An illustration of the graph comparing the actual data with the prediction results can be seen in Figure 4.

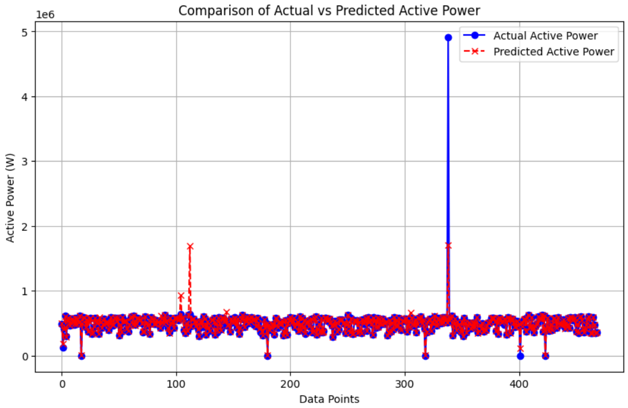


Fig. 4. Comparison of actual active power with model prediction calculations

Based on Figure 3, the majority of prediction results are consistent with the actual power level. However, at the data point around 300, there is a significant deviation in the prediction results. This variance is due to the presence of outliers, which are marked by extreme values in the pattern sequence, making it difficult for the model to capture and understand the pattern, resulting in a decrease in the concentration of prediction results. Linearly, outliers are also detected in the scatter plot graph, indicating a significant deviation in the prediction results. In the context of improving model performance, outliers in the form of extreme points need to be removed, but they also need to be retained to improve model generalization by considering prediction accuracy for unseen data [10]. The visualization of the scatter plot between actual active power and active power prediction can be described in Figure 5.

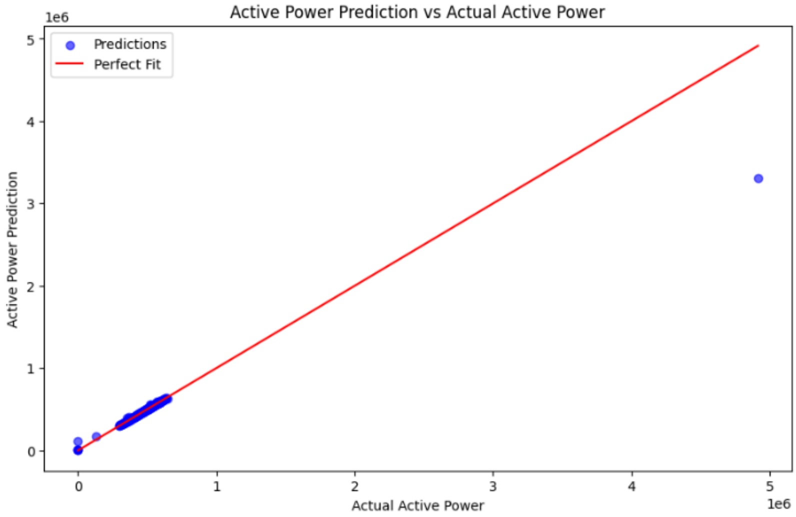


Fig. 5. Scatter plot between actual active power and predicted active power

4 Conclusion

This study describes the application of an optimized Random Forest algorithm to predict active power in electrical power systems based on voltage and current parameters. The model was evaluated using various performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R², which were used to assess the accuracy of the regression model.

The results show that the optimized Random Forest model achieved an RMSE of 74.308,10 W, an MAE of 5.911,61 W, and an R² of 0,8927, indicating that the model can explain approximately 89% of the variation in active power data. This indicates that the model can predict active power well, with most of the variation being explained by the features used.

However, the analysis also shows the presence of outliers, especially at several data points, where the predicted values deviate significantly from the actual values. This deviation is visible in the graph of the prediction results with Actual Active Power, which shows several data points where the model fails to provide accurate predictions for extreme values. This discrepancy can be attributed to the inherent variability in the dataset, especially in values outside the typical active power range.

Further research could focus on handling these outliers through data preprocessing techniques, such as outlier detection and removal, as well as exploring the use of other machine learning models or hybrid approaches to improve the model's robustness to extreme values. In addition, integrating temporal features or environmental factors could improve the model's predictive power for real-time dynamic forecasting. However, the existence of outliers also needs to be maintained in order to see data that has never appeared or been seen before.

References

[1] R. S. Kumar, P. S. Meera, V. Lavanya, and S. Hemamalini, “Brown bear optimized random forest model for short term solar power forecasting,” *Results Eng.*, vol. 25, p. 104583, 2025, doi: https://doi.org/10.1016/j.rineng.2025.104583.

[2] T. Ait tchakoucht, B. Elkari, Y. Chaibi, and T. Kousksou, “Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems,” *Energy Reports*, vol. 12, pp. 988–999, 2024, doi: https://doi.org/10.1016/j.egyr.2024.07.002.

[3] Z. Mustaffa and M. H. Sulaiman, “Random forest based wind power prediction method for sustainable energy system,” *Clean. Energy Syst.*, vol. 12, p. 100210, 2025, doi: https://doi.org/10.1016/j.cles.2025.100210.

[4] N. Mohapatra, K. Shreya, and A. Chinmay, “Optimization of the Random Forest Algorithm BT - Advances in Data Science and Management,” 2020, pp. 201–208.

[5] J. Gao, J. Ren, and Z. Wen, “Research on Diabetes Prediction Based on the Randomized Search CV Method,” in *2025 2nd International Conference on Electronic Engineering and Information Systems (EEISS)*, 2025, pp. 1–4. doi: 10.1109/EEISS65394.2025.11086023.

[6] D. Sun, H. Wen, D. Wang, and J. Xu, “A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm,” *Geomorphology*, vol. 362, p. 107201, 2020, doi: https://doi.org/10.1016/j.geomorph.2020.107201.

[7] Hanafi, “Detecting Hoax News Using Random Forest Algorithm with Hyperparameter Tuning GridSearchCV,” in *2025 4th International Conference on Electronics Representation and Algorithm (ICERA)*, 2025, pp. 683–688. doi: 10.1109/ICERA66156.2025.11087278.

[8] K. Chen, Y. Weng, A. A. Hosseini, T. Dening, G. Zuo, and Y. Zhang, “A comparative study of GNN and MLP based machine learning for the diagnosis of Alzheimer ’ s Disease involving data synthesis,” *Neural Networks*, vol. 169, no. May 2023, pp. 442–452, 2024, doi: 10.1016/j.neunet.2023.10.040.

[9] M. Ćalasan, S. H. E. A. 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 Convers. Manag.*, vol. 210, p. 112716, 2020, doi: 10.1016/j.enconman.2020.112716.

[10] P. K. Dongre, V. Patel, U. Bhoi, and N. N. Maltare, “An outlier detection framework for Air Quality Index prediction using linear and ensemble models,” *Decis. Anal. J.*, vol. 14, p. 100546, 2025, doi: https://doi.org/10.1016/j.dajour.2025.100546.