OPTIMIZATION OF HOUSEHOLD ENERGY USE PREDICTION USING RANDOM FOREST WITH GENETIC ALGORITHM FEATURE SELECTION

Nyimas Nisrinaa Kamilah [1], Abdul Rahman[2]
[1]Informatics, Faculty of Computer Science and Engineering, Universitas Multi Data Palembang
[2]Electrical Engineering, Faculty of Computer Science and Engineering, Universitas Multi Data Palembang
Road. Rajawali No.14, 9 Ilir, Kec. Ilir Tim. II, Palembang City, South Sumatra 30113
Email: nyimasnisrinaakamilah_2226250064@mhs.mdp.ac.id[1], arahman@mdp.ac.id[2]

Abstract - Electrical energy consumption continues to increase every year, so accurate prediction models are needed to support household electrical energy efficiency. This study analyzed high-resolution household electricity consumption data using the Random Forest (RF) algorithm and evaluated the influence of feature selection based on Genetic Algorithms (GA) in improving the performance of RF predictions. The base RF model achieves an RMSE of 0.6148, a MAE of 0.3478, and an R^2 of 0.5047. After implementing GA-based feature selection, the RF model with GA yields an RMSE of 0.6125 and an R^2 of 0.5084, indicating a marginal performance improvement. However, the MAE value increased slightly to 0.3503, which suggests that the increase was not uniform across the evaluation metrics. Overall, the RF approach with GA offers a modest improvement in prediction stability but with very limited accuracy gains, which highlights its potential and limitations in optimizing household energy consumption prediction.

Keywords - Feature Selection, Genetic Algorithm, Optimization, Prediction, Random Forest

I. INTRODUCTION

Electricity consumption continues to increase almost every year, although it had declined during the COVID-19 pandemic [1]. In Indonesia, the growth of electricity consumption is influenced by economic developments, an increase in the population, as well as an increase in the use of electronic devices in the household and industrial sectors [2], [3]. This condition not only adds to the burden on the energy supply system, but also has an impact on the availability of electricity and increases carbon [4], [5]. Badan Pusat Statistika (BPS) it also shows that the number and use of electronic devices continue to increase, and, without proper management, can lead to energy waste [6], [7]. Therefore, the community needs knowledge and tools to monitor and control energy use to support efficiency [8], [9]

The *Machine Learning* (ML) method is widely used in energy consumption prediction, one of which is Random Forest (RF), which is known to be stable and accurate compared to conventional approaches [10]–[12]. But, RF performance may decrease if there are irrelevant features in the data. To overcome this, RF is often combined with Genetic Algorithm (GA) as a feature selection method, where GA has been shown to improve model performance in a number of energy prediction studies [13]–[15]. Although the combination of RF and GA is not new, its application specifically to household energy consumption prediction is still limited, even though consumption patterns in this sector are very dynamic and are influenced by user behavior [16], [17]. Thus, the novelty of this research does not lie in the combination of algorithms, but in the context of its application, namely the application of RF with GA for high-resolution household consumption data and the analysis of the contribution of household appliance features. To identify research positions in the existing literature landscape, Bibliometric mapping was carried out based on keywords that often appear in publications related to energy consumption prediction. The results of the mapping are presented in Figure 1 as *a Bibliometric Network*.

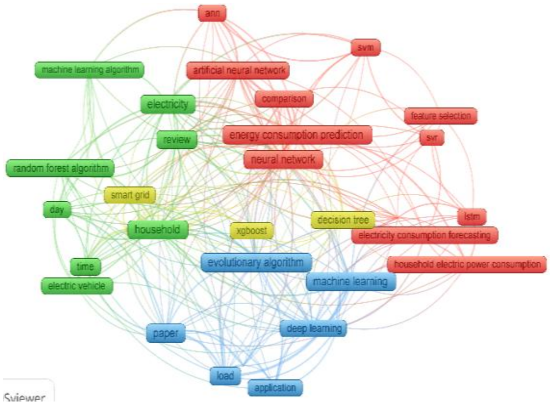

Figure 1. Bibliometric Network

Figure 1 shows a visualization of the co-occurrence of keywords that describes the structure of the research theme in the field of energy consumption prediction. Red clusters focus on neural network-based methods such as ANN, LSTM, and SVM, Blue clusters include machine learning, deep learning, and evolutionary algorithms, green clusters are related to energy domains such as household electricity consumption, smart grids, and RF algorithms, while the yellow cluster represents decision tree-based approaches such as Decision Tree and XGBoost. Connections between points indicate thematic relationships and the frequency of co-occurrence between keywords. But, the mapping also shows that although RF (in the green cluster) and GA (in the blue cluster) are both widely discussed, the combination of the two in the context of household energy prediction, especially with the features of home appliances and high data resolution, does not appear to be the dominant topic. This shows that there are research gaps that have not been explored much, i.e., the application of the RF hybrid method with GA, which is specifically aimed at understanding the contribution of equipment features in household energy consumption.

The bibliometric findings became the basis for the selection of methodology in this study. RF is used as the primary model due to its ability to handle non-linear energy consumption patterns, while GA is used to address RF weaknesses in the face of irrelevant features. The research focus on high-resolution household data was chosen to fill the research gaps that have been identified from bibliometric analysis. Thus, this study aims to analyze household electrical energy consumption using RF, evaluate the effectiveness of GA-based feature selection in improving model performance, and understand the features of household appliances that contribute the most to variations in energy consumption.

II. STUDY SIGNIFICANCE

A. Research Framework

This study is designed to develop a model for predicting household electrical energy consumption. This research requires a research flow and several stages. The method used to develop a prediction model for household electrical energy consumption is a hybrid approach of Random Forest with Genetic Algorithm. This hybrid approach was chosen because household electrical energy data is generally non-linear and complex, so it requires an algorithm that is able to handle many variables and feature interactions. The flow of this research has several stages, namely Data Collection, Data

Pre-Processing, Baseline Model, Feature Selection, Model Optimization, Model Evaluation, Analysis, and Conclusion. The research flow in this study can be seen in Figure 2.

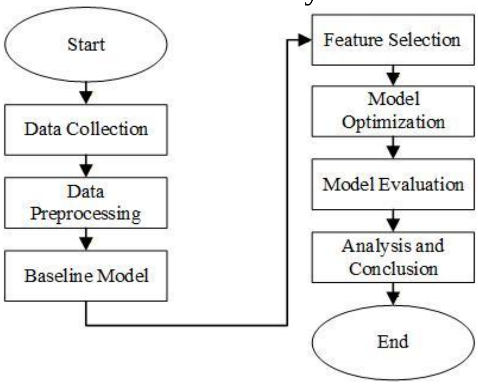

Figure 2. Research Framework

B. Data Collection

At this stage, the process of collecting high-resolution household electrical energy consumption datasets is carried out. The dataset is obtained in the form of multiple Excel sheets (.xlsx), where each sheet represents one house with a total of 35.137 rows and 13 columns. The number of rows shows that the data was collected in 15minute intervals over the course of an entire year, not one day, because 96 intervals per day × 365 days = 35.000 rows per home. Each sheet has the same feature format, namely: Unnamed:0, Periods, Total Consumption, AC, Dishwasher, Washing Machine, Dryer, Water Heater, TV, Microwave, Kettle, Lighting, *and* Refrigerator obtained from Zenodo [18]. To combine all sheets into one dataset, a vertical concatenation process is performed after adding House_ID columns as a marker of home identity. This process produces a total of 1.756.800 lines of data, which is 35.136 lines × 50 houses. All data is then used in the next pre-processing and modeling stage. An example of a dataset before being combined through *the vertical concatenation* process can be seen in Table 1.

Table 1. Data Collection

	Periods	**Total Consumption**	**AC**	**Dishwasher**	**Washing Machine**	**Dryer**	**Water heater**	**TV**	**Microwave**	**Kettle**	**Lighting**	**Refrigerator**
1	1	2,322958984	0	0	0	0	0	0	0	0	0	0
.	.	………	..	..	..	..	..	..	..	..	..	..
35.135	0	0	0	0	0	0	0	0	0	0	0	0

C. Data Preprocessing

The pre-processing stage of the data is carried out to ensure that the dataset is in a clean, consistent condition and ready to be used in the modeling process. Initial checks are performed to detect duplication, data type inconsistencies, and illogical consumption values. The results of the examination showed that there were missing values with a very small proportion, that is, less than 0.5% in some numeric columns. All missing values are then visualized using a missing values map to verify their distribution. Given the low proportions and distribution of features that do not show extreme skewness, imputation is carried out using the mean value. The choice of the mean method was considered based on three reasons:

1. The proportion of missing values is very small so the risk of bias is minimal.
2. The mean maintains consistency of the energy data scale between equipment.
3. mean is appropriate for relatively stable distributed features on short-interval energy datasets.

Furthermore, feature transformations are carried out, including the creation of time-based features such as Periods to capture daily consumption patterns. The features of household appliances such as air conditioners, water heaters, lighting, and refrigerators are then normalized using Min–Max Scaling to standardize the range of values between features that have different scales. After this process, all the data is in a consistent numerical format and free of missing values. The dataset is then divided into 80% training data and 20% test data, by ensuring that the distribution of each feature remains stable across both subsets of data. The complete pre-processing stage can be seen in Figure 3.

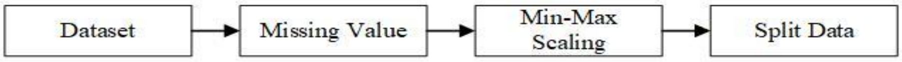

Figure 3. Preprocessing Stages

D. Baseline Model

This stage is the process of building a baseline model using the Random Forest algorithm as a performance reference before optimizing using the Genetic Algorithm. This basic model is trained using all features of pre-processing results without feature selection, thus providing an overview of the model's initial ability to predict household energy consumption. The selection of Random Forest was based on its stability in handling high-dimensional data as well as its ability to model non-linear relationships, so it is widely recommended in machine learning-based energy prediction studies. At this stage, the dataset is divided into 80% training data and 20% test data, with random partitions replicated using fixed random_state.

The Random Forest model was built using a consistent configuration throughout the experiment, namely with a number of trees (n_estimators) as many as 200, a random_state value of 42, a loss function in the form of squared_error, and a bootstrap setting that remains activated with the selection of maximum features following the auto value. All normalized numerical features are used as model inputs. The results of this model baseline training are then used as a benchmark to evaluate performance improvement after the application of the Genetic Algorithm-based feature selection method. The stages of development of the Baseline Model are shown in Figure 4.

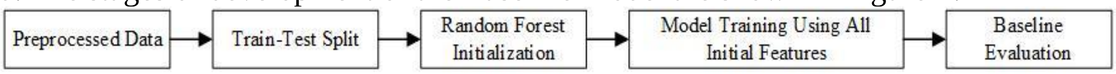

Figure 4. Baseline Model Stage

E. Feature Selection

The feature selection stage using Genetic Algorithm (GA) is carried out to obtain a combination of features that provides the best prediction performance. At this stage, each chromosome is represented as a binary vector indicating whether a feature is selected (1) or not selected (0). The process of population initialization is carried out by generating a random number of chromosomes, then each chromosome is evaluated through a fitness value calculated based on the performance of the Random Forest model trained using a subset of the features in question. This approach is in line with common practice in predictive model optimization research based on evolutionary algorithms, which utilize *the ensemble* model as a fitness evaluator.

In each generation, GA runs a selection stage to select the best chromosome, followed by a crossover process to produce a combination of new features, as well as mutations to maintain the diversity of solutions. This iterative process is carried out over a number of generations so that features that have a large contribution to the model's accuracy have a greater chance of being retained and combined in subsequent populations. The final result of this stage is a subset of optimal features that are used as input for the Random Forest model at the stage of building the Random Forest model with Genetic Algorithms. The flow of the Feature Selection stages can be seen in Figure 5, and the dataset after feature selection can be seen in Table 2.

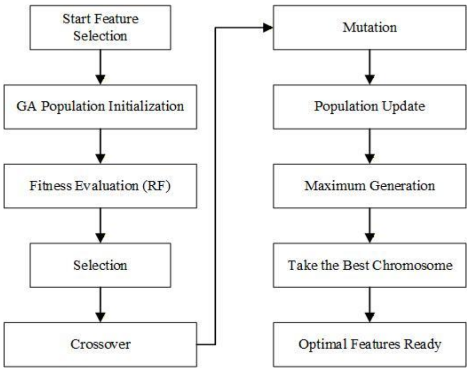

Figure 5. Feature Selection

Table 2. Dataset After Feature Selection

	Unnamed: 0	Periods	AC	Dishwasher	Washing Machine	Dryer	Water heater	TV	Microwave	Kettle	Lighting
1106456	17240	57	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000	0.0	0.0
115280	9872	81	0.0	0.0	0.000000	0.0	0.0	0.152325	0.000	0.0	0.0
486419	29651	84	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000	0.0	0.0

F. Model Optimization

The model optimization stage is carried out by running the Genetic Algorithm iteratively to obtain a subset of the best features used in the construction of the final Random Forest model. The optimization process lasts for generations, where in each generation an evaluation of all chromosomes is carried out based on fitness values calculated through the performance of the Random Forest model on a subset of features represented by the chromosomes.

During the evolutionary process, the best-performing chromosomes are preserved through the mechanism of selection, while new chromosomes are generated through crossover and mutation operations to explore possible combinations of more optimal features. When fitness values do not show significant improvement or when the maximum number of generations is reached, GA generates a subset of optimal features that are then used to build optimized Random Forest models (RF with GA). This RF with GA model utilizes only the most relevant features so that it is more computationally efficient and focuses on the variables that contribute the most to predictive performance.

G. Model Evaluation

This stage is an evaluation process to compare the performance of the baseline model and the RF optimization result model with GA. The evaluation was carried out using three main metrics, namely Root Mean Square Error (RMSE) in equation (1), Mean Absolute Error (MAE) in equation (2), and determination coefficient R^2 in equation (3). These three metrics are widely used in household energy consumption prediction research, especially in studies that apply ensemble and hybrid methods [19], so it is relevant to measure the accuracy and stability of the model.

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

Where y_i is the actual value, $\hat{y}_i$ is the model's predictive value, n is the sum of data, and $\bar{y}$ is the average of the actual value. In this study, the evaluation was carried out using 20% of the test data that had been separated at the pre-processing stage. The validation strategy using a one-time 80% training data and 20% test data was chosen due to the very large size of the dataset (1.7 million lines). The use of other validation methods, such as k-fold cross-validation, has the potential to incur a much higher and inefficient computational load for this dataset size. In addition, data sharing is carried out randomly with the determination of random_state to maintain the reproducibility of the results.

RMSE and MAE values were used to measure the magnitude of the average model prediction error, while R^2 values were used to measure the model's ability to explain the variability of energy consumption in the dataset. By comparing the three metrics directly between the baseline model and the RF–GA model, the extent to which the GA-based feature selection process contributes to improving the accuracy, efficiency, and stability of the prediction model can be assessed.

H. Analysis and Conclusion

The analysis stage in this study was carried out by comparing two modeling approaches, namely the Random Forest baseline model, which is built using all features, and the Random Forest model, which has gone through a feature selection process using Genetic Algorithm (GA). Analysis was carried out on the value of evaluation metrics such as RMSE, MAE, and R^2 to assess the influence of feature selection on model performance. This process allows the identification of differences in model characteristics before and after feature optimization, thus providing an overview of the contribution of each feature and the effectiveness of the GA approach in the modeling process.

Moreover, the analysis also includes observations of the features selected by the GA to understand which variables are considered most relevant in influencing energy consumption. This stage is structured to provide a solid basis for the interpretation of the optimization process and ensure that the modeling steps used are aligned with the principles of data analysis and machine learning. The conclusions at this stage do not refer to the final results of the study, but describe how the analysis process is carried out in the research to evaluate the model methodologically.

III. RESULTS AND DISCUSSION

In this study, a number of experimental stages were carried out to obtain predicted results of household electrical energy consumption using the Random Forest model and the Random Forest model optimized with Genetic Algorithm (GA). The results presented in this section are in the form of model evaluation values, features selected by GA, and performance comparison between the baseline model and the optimization result model. The presentation of the results is carried out without further interpretation.

A. Preprocessing Results

In the preprocessing stage, data quality checks are carried out by visualizing the presence of missing values, as shown in Figure 6, which ensures that all features have no lost values so that the dataset can be used directly in the modeling process. Next, the distribution of data before and after normalization is shown in Figure 7, where normalization using MinMaxScaler makes the values on each feature more uniform, thereby increasing stability at the model training stage. Moreover, Figure 8 shows a numerical feature correlation heatmap used to observe the relationships between variables and identify initial correlation patterns before further feature selection and modeling processes.

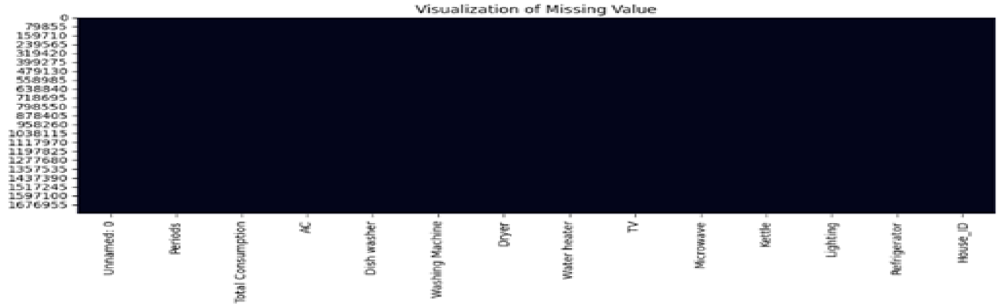

Figure 6. Visualization of Missing Value

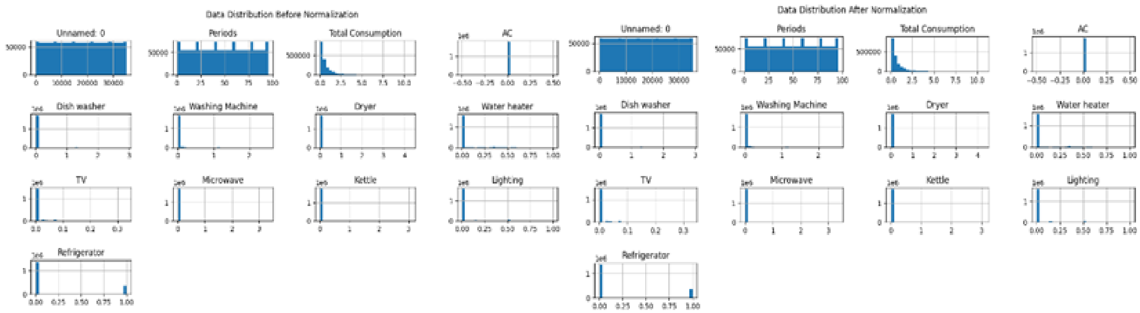

Figure 7. Data Distribution Before and After Normalization

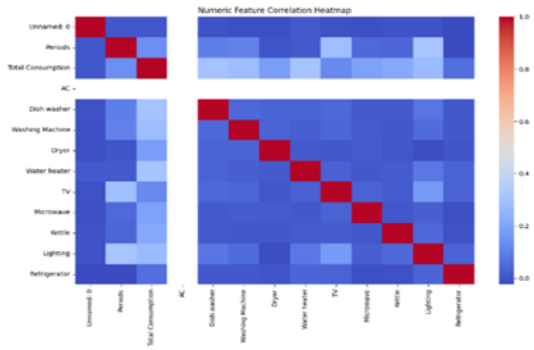

Figure 8. Numeric Feature Correlation Heatmap

B. Random Forest Baseline Model Results

The baseline model uses the Random Forest Regressor algorithm with 200 trees without feature optimization, as seen in Figure 9. Based on the predictions in the test data, the evaluation value can be seen in Table 3.

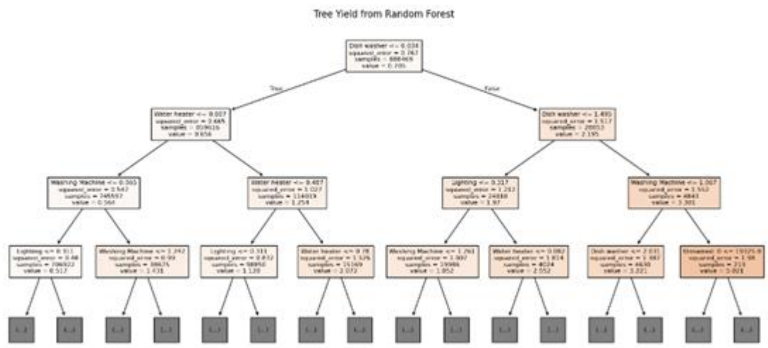

Figure 9. Tree Yield from Random Forest

Table 3. Random Forest Results

Metric	Value
RMSE	0.6148
MAE	0.3478
R^2	0.5047

An R^2 value of 0.5047 indicates that the model is only able to explain about 50% of household energy consumption variance, So that half of the pattern of consumption variation was not successfully predicted by the model. This condition indicates that the variables used in the model that only include the features of household appliances are not yet fully able to capture the complexity of energy consumption at 15-minute intervals. The high volatility of household energy consumption, which is influenced by user behavior, weather conditions, number of residents, and daily activity patterns, has the potential to be a factor that is not captured by the model. In addition, RF, as a tree-based model, has limitations in modeling temporal dynamics without richer time features. Thus, the baseline results of this model give an idea that although the Random Forest is quite stable, the model still has limitations in explaining energy consumption patterns as a whole, so that a feature optimization process or the addition of supporting variables is needed to improve its performance.

C. Genetic Algorithm Feature Selection Results

The optimization process was carried out using GA with an initial configuration in the form of a population of 51 individuals that was evolved over 10 generations. Individual selection was carried out using the tournament selection method, while the reproduction process used two-point crossover. The mutation operator used is flip-bit with a mutation probability of 0.1. GA runs for 10 generations and shows a gradual increase in R^2 value from 0.4133 to 0.4327 in the 10th generation. GA then generates a subset of 10 optimal features, namely, Unnamed: 0, Periods, Dishwasher, Washing Machine, Dryer, Water heater, TV, Microwave, Kettle, Lighting.

These features were chosen as the most relevant variables in explaining household energy consumption based on the fitness function R^2.

D. Random Forest Results After Genetic Algorithm Optimization

The Random Forest model was then retrained using 10 selected features through a feature selection process based on GA. The results of the model evaluation after optimization are shown in Table 4.

Table 4. Random Forest Results After GA Optimization

Metric	**Value**
RMSE	0.6125
MAE	0.3503
R^2	0.5084

Compared to *the baseline model*, it can be seen that the performance improvement obtained is relatively small. The R^2 value only increased from 0.5047 to 0.5084, indicating that GA's contribution to the explanation of energy consumption variance only increased by about 0.37%, which is practically insignificant. The RMSE decrease is also only 0.0023, so it does not reflect a significant increase in accuracy. In fact, the MAE value has increased slightly from 0.3478 to 0.3503, which indicates that the absolute error average is actually increasing. These results show that even though GA is able to select a more compact subset of features, the GA configuration with 10 generations and a population of 51 is still too limited to optimally explore the solution space. As a result, the improvements achieved are more cosmetic than substantial performance improvements. This is consistent with the characteristics of fluctuating household energy consumption data and is strongly influenced by external factors not available in the dataset.

E. Model Comparison

A comparison of the baseline model and the model after GA optimization can be seen in Table 5.

Table 5. Model Comparison

Metric	**RF**	**RF + GA**	**Change**
RMSE	0.6148	0.6125	Better
MAE	0.3478	0.3503	Slightly up
R^2	0.5047	0.5084	Increase

Based on the table, it can be seen that the performance improvement after GA implementation is very minimal. The RMSE decrease was only 0.0023, while the R^2 increase was only about 0.37%, which does not show a significant improvement in the model's ability to explain the data variance. The increase in MAE shows that for some data points, the prediction has become less accurate. These findings indicate that the feature selection process by GA in the configured used has not been able to provide a significant increase in accuracy. The limitation of the number of generations and the size of the population causes the space to search for solutions not to be explored optimally. In addition, the highly volatile nature of household energy consumption data causes feature optimization to have a significant impact on model performance.

This section discusses the main findings of the study based on the performance of the baseline Random Forest model and the Random Forest model optimized with Genetic Algorithm (RF–GA). Evaluations were carried out using RMSE, MAE, and R^2. The baseline results showed that Random Forest produced an RMSE of 0.6148, MAE of 0.3478, and R^2 of 0.5047. The baseline results showed that Random Forest produced an RMSE of 0.6148, MAE of 0.3478, and R^2 of 0.5047. The R^2 value, which is only able to explain about 50% of the variability in energy consumption, indicates that many external factors, such as the number of inhabitants, weather, activity patterns, and user behavior, are not covered by the features of household appliances. This is in accordance with the characteristics of short-interval household energy consumption, which is highly volatile and difficult to accurately predict using tree-based models[20].

After the feature selection using GA, there was a very small change in performance; the RMSE decreased slightly to 0.6125, and the R^2 increased to 0.5084, but the MAE increased to 0.3503. These improvements are minimal and not practically significant, so GA's contribution in the configuration of this study is more about minor stability, not increased accuracy. Selected features such as Dishwasher, Washing Machine, Dryer, Water Heater, Lighting, and Refrigerator show that appliances with large power consumption play an important role in short-term energy consumption patterns. However, the effectiveness of GA is limited due to the low evolutionary parameters of only 10 generations with a population of 51, so the space for exploration of solutions is not optimal. The literature also shows that GA with minimal generation tends to stop at sub-optimal solutions [21].

Although the improvement is small, the feature selection process through GA is still beneficial in terms of computational efficiency and interpretability. The reduction in the number of features makes the model lighter to process on large datasets (~1.7 million rows) without drastically degrading performance. In addition, the selected features provide an overview of which devices have the most effect on short-interval energy consumption. Overall, this study shows that the Random Forest is a stable model but has limitations in capturing highly dynamic energy consumption variability. GA makes a positive but limited contribution, mainly due to the minimal evolutionary configuration and absence of external features. GA makes a positive but limited contribution, mainly due to the minimal evolutionary configuration and absence of external features.

IV. CONCLUSION

This study evaluates the ability of RF and feature selection optimization using GA in predicting household energy consumption. The results showed that the baseline RF model was only able to explain about 50% of the data variance ($R^2 = 0.5047$), so the performance is still limited to high-resolution energy data. The implementation of GA results in a very small increase ($R^2 = 0.5083$), so its contribution to model accuracy can be said to be minimal and has not shown significant optimization. The main conclusion of this study is that the RF–GA combination can reduce the number of features and provide a slight improvement in model stability. However, it is not enough to produce a significant increase in accuracy on this dataset. These results also indicate that the highly volatile character of the 15-minute interval energy consumption, as well as the GA configuration with a limited number of generations, are factors limiting the performance of the model.

This research has several limitations, namely the use of one dataset, one main algorithm, and the absence of cross-validation, and does not include external variables such as weather or the number of inhabitants. For further research, it is recommended the use a larger number of GA generations, exploration of other optimization algorithms (e.g., PSO, Jaya, or ALO), as well as the addition of contextual features to improve the generalization capabilities of the model.

REFERENSI

[1] A. E. Septiyono, *Statistik Listrik Electric Statistics 2019-2023*. Jakarta: Badan Pusat Statistik, 2024.

[2] A. A. Solikah and B. Bramastia, "Systematic Literature Review : Kajian Potensi dan Pemanfaatan Sumber Daya Energi Baru dan Terbarukan Di Indonesia," *J. Energi Baru dan Terbarukan*, vol. 5, no. 1, pp. 27–43, 2024, doi: 10.14710/jebt.2024.21742.

[3] I. F. U. Muzayanah, H. H. Lean, D. Hartono, K. D. Indraswari, and R. Partama, "Population density and energy consumption: A study in Indonesian provinces," *Heliyon*, vol. 8, no. 9, p. e10634, 2022, doi: 10.1016/j.heliyon.2022.e10634.

[4] D. Kamani and M. M. Ardehali, "Long-term forecast of electrical energy consumption with considerations for solar and wind energy sources," *Energy*, vol. 268, p. 126617, 2023, doi: 10.1016/j.energy.2023.126617.

[5] H. Dissanayake *et al.*, "Nexus between carbon emissions, energy consumption, and economic growth: Evidence from global economies," *PLoS One*, vol. 18, no. 6 June, pp. 1–27, 2023, doi: 10.1371/journal.pone.0287579.

[6] Badan Pusat Statistika, "Ekonomi Indnesia Triwulan III-2022 Tumbuh 5,72 persen (y-on-y)," *Www.Bps.Go.Id*, no. 13, p. 12, 2022, [Online]. Available: https://www.bps.go.id/id/pressrelease/2022/11/07/1914/ekonomi-indonesia-triwulan-iii-2022-tumbuh-5-72-persen--y-on-y-.html.

[7] T. H. Tran and T. B. T. Nguyen, "Minimizing Total Cost of Home Energy Consumption under Uncertainties," *Electr. Power Components Syst.*, vol. 50, no. 19–20, pp. 1143–1160, 2022, doi: 10.1080/15325008.2022.2148774.

[8] H. A. Yuniarto *et al.*, "Customer's response to dynamic pricing in utility energy Tariff quality and reliability with the time of use: An Empirical case study of household electricity customers in Indonesia," *Heliyon*, vol. 10, no. 11, p. e31415, 2024, doi: 10.1016/j.heliyon.2024.e31415.

[9] G. Sart, M. H. Ozkaya, and Y. Bayar, "Education, Financial Development, and Primary Energy Consumption: An Empirical Analysis for BRICS Economies," *Sustain.*, vol. 14, no. 12, 2022, doi: 10.3390/su14127377.

[10] F. Dinmohammadi, Y. Han, and M. Shafiee, "Predicting Energy Consumption in Residential Buildings Using Advanced Machine Learning Algorithms," *Energies*, vol. 16, no. 9, 2023, doi:

10.3390/en16093748.

[11] Y. Retnowati, A. Mafahim, and I. Syawanodya, “Comparative Analysis of Artificial Neural Network (ANN) and Random Forest in House Price Prediction,” *J. Softw. Eng. Inf. Commun. Technol.*, vol. 5, no. 2, pp. 135–150, 2024.

[12] S. Sartini, L. Rohimah, Y. I. Maulana, S. Supriatin, and D. Yuliandari, “Optimization of Random Forest Prediction for Industrial Energy Consumption Using Genetic Algorithms,” *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 11, no. 1, pp. 35–44, 2023, doi: 10.33558/piksel.v11i1.5886.

[13] M. Dostmohammadi, M. Z. Pedram, S. Hoseinzadeh, and D. A. Garcia, “A GA-stacking ensemble approach for forecasting energy consumption in a smart household: A comparative study of ensemble methods,” *J. Environ. Manage.*, vol. 364, no. May, p. 121264, 2024, doi: 10.1016/j.jenvman.2024.121264.

[14] A. G. Putrada, N. Alamsyah, I. D. Oktaviani, and M. N. Fauzan, “A Hybrid Genetic Algorithm-Random Forest Regression Method for Optimum Driver Selection in Online Food Delivery,” *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 4, pp. 1060–1079, 2023, doi: 10.26555/jiteki.v9i4.27014.

[15] R. Shintabella, C. E. Widodo, and A. Wibowo, “A Hybrid Machine Learning-Genetic Algorithm (ML-GA) to Predict Power Tranformer Life,” *Int. J. Heal. Educ. Soc.*, vol. 6, no. 12, pp. 26–38, 2023.

[16] S. Piu, A. Arifin, and M. Rizal, “Optimization of Electric Energy Usage for Household Customers Based on Machine Learning and Internet of Things Optimasi,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 1, pp. 84–92, 2024, doi: 10.57152/malcom.v5i1.1694.

[17] F. Prayoga, Najwa Aulia Meliala, Dhea Aulia, Seroja Fi Maaris Mar’ah, Harly Bima Perkasa Alam, and Mhd. Furqan, “Utilization of Data Mining in Finding Household Electricity Consumption Patterns in Indonesia,” *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 3, pp. 2549–2552, 2025, doi: 10.59934/jaiea.v4i3.1206.

[18] C. Goncalves, R. Barreto, P. Faria, L. Gomes, and Z. Vale, “Dataset of an energy community’s generation and consumption with appliance allocation,” *Data Br.*, vol. 45, 2022, doi: 10.1016/j.dib.2022.108590.

[19] G. Airlangga, “Evaluation of Ensemble and Hybrid Models for Predicting Household Energy Consumption : A Comparative Study of Machine Learning Approaches,” *Indones. J. Artif. Intell. Data Min.*, vol. 8, no. 1, pp. 94–101, 2025, doi: 10.24014/ijaidm.v8i1.32819.

[20] C. B. Pop, V. R. Chifu, C. Cordea, E. S. Chifu, and O. Barsan, “Forecasting the Short-Term Energy Consumption Using Random Forests and Gradient Boosting,” *IEEE*, pp. 1–6, 2021, doi: 10.1109/RoEduNet54112.2021.9638276.

[21] J. Ding, J. Du, H. Wang, and S. Xiao, “OPEN A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning,” *Sci. Rep.*, vol. 15, p. 16828, 2025, doi: 10.1038/s41598-025-01761-1.

ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to Multi Data University of Palembang, especially the Faculty of Computer Science and Engineering and the Informatics Study Program, for the facilities, support, and guidance provided during the implementation of this research. Thanks are also extended to all parties who have assisted in the data collection, programming, and research analysis. Finally, the author expresses his appreciation to the reviewers and editors for the constructive input that has improved the quality of this manuscript.