

ELECTRICITY CONSUMPTION PREDICTION AND INFLUENTIAL FACTORS ANALYSIS USING MACHINE LEARNING REGRESSION

Evita Fitri^{1*}, Siti Nurhasanah Nugraha¹, Muji Ernawati²

Information System¹
Informatics²

Universitas Nusa Mandiri, Jakarta, Indonesia^{1,2}

<https://www.nusamandiri.ac.id/>^{1,2}

evita.etv@nusamandiri.ac.id*, siti.nhg@nusamandiri.ac.id, muji.mei@nusamandiri.ac.id

(*) Corresponding Author

(Responsible for the Quality of Paper Content)



The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract— The increase in electricity demand in line with population growth and economic activity requires an accurate and reliable electricity consumption forecasting system. Short-term electricity consumption predictions are an important component in energy system planning and management, particularly to support grid stability and operational efficiency. This study aims to model electricity consumption predictions using a machine learning regression approach and analyze the factors that most influence electricity consumption based on historical data. The dataset used consists of smart meter data with a 30-minute time interval that has undergone data cleansing, data transformation, and feature engineering, including the formation of lag features and temporal features. Three regression algorithms were used, namely Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression. Model evaluation was performed using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) metrics. The results show that Linear Regression provides the best performance on the test data with an RMSE value of 0.156, MAE of 0.125, and R^2 of 0.140, and demonstrates stable generalization capabilities. The analysis of influencing factors reveals that historical consumption variables, particularly Avg_Past_Consumption and electricity consumption lag features, are dominant factors in the prediction, while environmental variables contribute relatively less. These findings provide practical implications for short-term energy demand planning by enabling more accurate load estimation and supporting data-driven decision-making through interpretable electricity consumption patterns.

Keywords: Electricity Consumption Prediction, Feature Engineering, Machine Learning Regression, Smart Meter Data, Time Series.

Intisari— Peningkatan kebutuhan energi listrik seiring pertumbuhan populasi dan aktivitas ekonomi menuntut adanya sistem peramalan konsumsi listrik yang akurat dan dapat diandalkan. Prediksi konsumsi listrik jangka pendek menjadi komponen penting dalam perencanaan dan pengelolaan sistem energi, khususnya untuk mendukung stabilitas jaringan dan efisiensi operasional. Penelitian ini bertujuan untuk memodelkan prediksi konsumsi listrik menggunakan pendekatan machine learning regression serta menganalisis faktor-faktor yang paling berpengaruh terhadap konsumsi listrik berbasis data historis. Dataset yang digunakan berupa data smart meter dengan interval waktu 30 menit yang telah melalui tahap data cleansing, transformasi data, dan feature engineering, termasuk pembentukan lag features dan fitur temporal. Tiga algoritma regresi digunakan, yaitu Linear Regression, Random Forest Regression, dan Gradient Boosted Trees Regression. Evaluasi model dilakukan menggunakan metrik Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan Coefficient of Determination (R^2). Hasil penelitian menunjukkan bahwa Linear Regression memberikan performa terbaik pada data pengujian dengan nilai RMSE sebesar 0.156, MAE sebesar 0.125, dan R^2 sebesar 0.140, serta menunjukkan kemampuan generalisasi yang stabil. Analisis faktor berpengaruh mengungkapkan bahwa variabel konsumsi historis, khususnya Avg_Past_Consumption dan fitur



lag konsumsi listrik, merupakan faktor dominan dalam prediksi, sementara variabel lingkungan memberikan kontribusi yang relatif lebih kecil. Temuan ini menunjukkan bahwa model regresi yang sederhana dan interpretable tetap relevan untuk prediksi konsumsi listrik jangka pendek, sekaligus memberikan pemahaman yang jelas mengenai faktor-faktor yang memengaruhi konsumsi energi listrik.

Kata Kunci: *Prediksi Konsumsi Listrik, Rekayasa Fitur, Regresi Pembelajaran Mesin, Data Meter Pintar, Deret Waktu.*

INTRODUCTION

Electricity consumption plays a crucial role in the planning and management of modern energy systems. The continuous increase in electricity demand, driven by population growth, urbanization, and economic activities, necessitates the development of efficient and sustainable energy management strategies [1]. Forecasting electricity consumption—particularly during peak load periods—is essential for planning energy production and distribution, as well as for reducing operational costs and environmental impacts. Previous studies have indicated that accurate forecasting approaches can help energy providers anticipate demand surges and mitigate the risk of supply disruptions [2]. Conversely, inaccurate forecasting may lead to various issues, including energy waste, load imbalances, and increased operational costs within the power system [3]. Therefore, accurate electricity consumption prediction is a critical requirement for energy providers and policymakers to support both short-term and long-term energy planning.

Several studies have highlighted that electricity consumption exhibits time-dependent characteristics, where the consumption level at a given period is influenced by patterns observed in previous periods [4], [5], [6]. In addition to historical consumption data, electricity consumption is also affected by external variables such as temperature, humidity, and environmental conditions [4]. These characteristics make electricity consumption forecasting a complex task that requires modeling approaches capable of capturing temporal patterns and the influence of multiple contextual variables effectively [7].

With the advancement of technology, machine learning approaches have increasingly been adopted for electricity consumption prediction. Various regression algorithms, including Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression, have been utilized to improve prediction accuracy by leveraging historical data and additional explanatory variable [8], [9]. Ensemble-based models have been reported to provide strong predictive performance in cases where the data

exhibit complex non-linear relationships [10]. Meanwhile, simpler models such as Linear Regression remain widely used due to their interpretability and transparency in explaining relationships among variables [11].

Despite these developments, previous studies have reported varying findings concerning the factors that most significantly influence electricity consumption. Some studies emphasize the importance of weather-related factors, while others demonstrate that electricity consumption is primarily driven by historical usage patterns and past load behavior [12], [13]. Moreover, many existing studies focus heavily on improving predictive accuracy, yet provide limited attention to model interpretability and comprehensive analysis of influential factors [14], [15], [16], [17], [18].

A recent study [19] proposed a deep learning-based two-phase framework for forecasting household energy consumption using Social IoT smart meter data. In the first phase, a Long Short-Term Memory (LSTM) model was employed to predict total active power for up to 500 hours ahead. In the second phase, a Convolutional LSTM (ConvLSTM) architecture was used to forecast weekly household energy consumption by integrating convolutional features with temporal learning. Experimental results showed that ConvLSTM achieved the lowest RMSE, demonstrating the effectiveness of deep learning in capturing complex, long-term consumption patterns. However, this approach primarily focused on predictive accuracy and did not provide interpretative insights into the factors influencing consumption, highlighting the need for simpler and more explainable modeling techniques.

A study from [20] showed that deep learning approaches, such as LSTM, can achieve very high levels of accuracy in forecasting electricity consumption when applied to large-scale and high-temporal resolution datasets. Nevertheless, deep learning models generally require substantial data volume, complex training processes, and exhibit limited interpretability—particularly in identifying explicit factors that influence electricity consumption. While these models demonstrate superior predictive capability, the use of simpler and more interpretable methods remains relevant

for understanding underlying consumption patterns. Therefore, this study explores lightweight and interpretable machine learning regression techniques and examines dominant factors influencing electricity consumption using historical data.

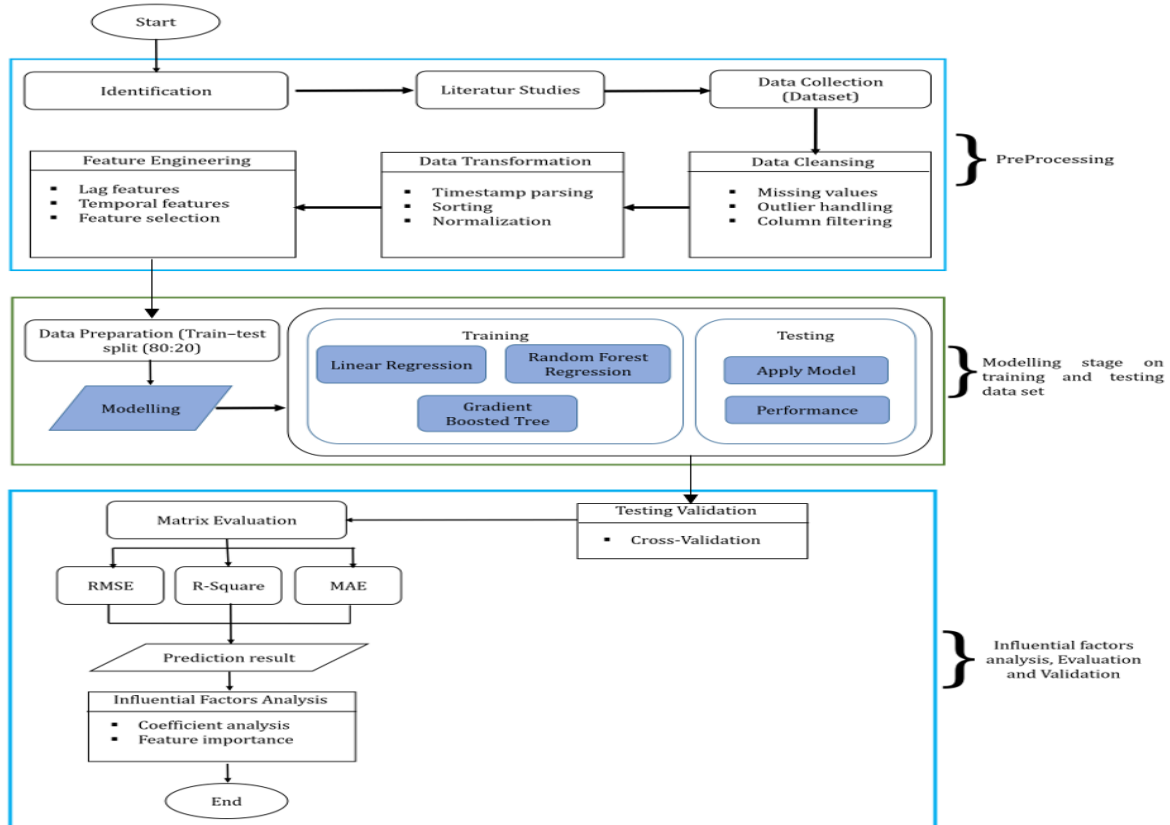
In addition, research by [21] demonstrated that electricity consumption prediction at high temporal resolution, such as 15-minute intervals, allows for improved forecasting accuracy and facilitates the identification of fine-grained consumption patterns. Although previous studies have successfully developed AI-based forecasting models using high-frequency data, most of these studies have primarily emphasized predictive performance and system implementation, while interpretability and influential factor analysis remain underexplored. Despite the identification of influencing variables in earlier work, the integration of electricity consumption forecasting with interpretable regression-based factor analysis is still limited.

Based on the identified problem, this study focuses on developing machine learning regression models for electricity consumption prediction while considering both accuracy and interpretability. This

study compares the performance of three regression models—Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression—using time-based electricity consumption data enhanced with feature engineering, including historical consumption (lag features) and temporal features. Furthermore, this study conducts an influential factors analysis to identify the most dominant variables affecting electricity consumption. The primary contribution of this work is the presentation of a comprehensive analysis that not only evaluates predictive accuracy but also provides a clear understanding of the factors influencing electricity consumption based on historical patterns.

MATERIALS AND METHODS

This research was conducted in several main stages, namely problem identification, literature study, data collection, and data pre-processing. The next stage was the design and training of regression models using three machine learning algorithms, followed by model evaluation and analysis of influencing factors. The overall research workflow is illustrated in Figure 1.



Source: (Research Results, 2025)

Figure 1. Research Stages



Based on Figure 1, the research method stages consist of several main steps, including:

Identification

This identification stage involves understanding the main issues related to electricity consumption prediction needs. At this stage, initial observations are also made on the phenomenon of increased energy consumption and the importance of electricity load prediction in power system planning. In addition, a preliminary review of the dataset and mapping of analysis needs are also carried out, including: Determine the focus of the research, namely short-term electricity consumption prediction. Set the target variable to be predicted (Electricity_Consumed). Determine the input variables that could potentially have an effect, including historical variables and environmental variables and Identify the appropriate machine learning method for analysis.

Literature Study

This stage is conducted to gain an in-depth understanding of:

1. Electricity consumption prediction techniques that have been widely used, such as regression, ensemble learning, and deep learning.
2. Factors that have been proven to affect electricity consumption, such as air temperature, humidity, and historical energy usage patterns.
3. Previous relevant research, including machine learning and feature engineering approaches to time series datasets.
4. The advantages and limitations of prediction models previously used by researchers.

Data Collection

At this stage, electricity consumption datasets are collected as the main material for the modeling process.

1. Data Source

The dataset used in this study was obtained from the Kaggle platform, titled "Smart Meter Electricity Consumption Dataset", which is publicly accessible at: <https://www.kaggle.com/datasets/ziya07/smart-meter-electricity-consumption-dataset>.

This dataset provides electricity consumption data obtained from smart meters and has been supplemented with weather condition variables and historical electricity consumption statistics.

2. Dataset Structure

The dataset has time series characteristics, meaning that data is collected at 30-minute intervals, so the relationship between time is very important in determining consumption patterns. This characteristic indicates that electricity consumption at a given time is influenced by previous consumption (time-dependent). The dataset in this study contains several variables, namely:

Table 1. Dataset Variable Details

No	Variable Name	Description
1	Timestamp	Measurement timestamp at 30-minute intervals.
2	Electricity_Consumed (kWh)	The amount of electricity consumption at each time interval. This variable is the prediction target.
3	Temperature (°C)	External environmental temperatures that can affect electricity usage.
4	Humidity (%)	Air humidity level
5	Wind_Speed (km/h)	Wind speed that could potentially affect the comfort of the place and electricity consumption.
6	Avg_Past_Consumption (kWh)	Average historical consumption over a specific period (rolling average).
7	Anomaly_Label	Categorical labels to mark normal or anomalous usage (results from the Isolation Forest model).

Source: (Research Results, 2025)

3. Data Characteristics

The dataset consists of 5000 rows of data recorded at 30-minute intervals. This amount of data is sufficient to form temporal patterns and build lag features at the feature engineering stage. There are anomaly labels, but this study focuses on predicting electricity consumption rather than anomaly classification. The following is an example of the dataset used in this study.

Table 2. Sample Dataset

No	Timestamp	Electricity_Consumed	Temperature	Humidity	Wind_Speed	Avg_Past_Consumption	Anomaly_Label
1	01/01/2024 00:00	0.45778569216 85388	0.46952445708 73399	0.39636835925 751607	0.44544059952 876924	0.69205721068 88903	Normal
2	01/01/2024 00:30	0.35195594980 48026	0.46554477464 769306	0.45118441315 07186	0.45872928645 142597	0.53987373576 85197	Normal
3	01/01/2024 01:00	0.48294808968 34488	0.28541498990 965564	0.40828896968 83288	0.47036041265 32033	0.61472396090 16935	Normal



No	Timestamp	Electricity_Consumed	Temperature	Humidity	Wind_Speed	Avg_Past_Consumption	Anomaly_Label
4	01/01/2024 01:30	0.62883830940 13377	0.48209508722 60694	0.51230839684 13018	0.57624138712 85659	0.75704435194 89565	Normal
5	01/01/2024 02:00	0.33597443754 611067	0.62474094778 27948	0.67202066798 6486	0.37300379071 0542	0.67398094942 99462	Normal
6	01/01/2024 02:30	0.33597717384 180326	0.35538938052 48106	0.38276742334 73963	0.36111126312 146946	0.61860665935 09736	Normal
7	01/01/2024 03:00	0.63820213591 78986	0.66710227864 02212	0.64316353896 34667	0.61100351124 50067	0.70322512861 45852	Normal
8	01/01/2024 03:30	0.50290578819 21515	0.59185418321 21293	0.38120523193 19284	0.40714378789 54785	0.71804984594 60256	Normal
9	01/01/2024 04:00	0.29675426901 084134	0.42785828294 43013	0.37151234616 993983	0.31112141456 52017	0.66370318620 59055	Normal
10	01/01/2024 04:30	0.46542667393 099413	0.50740028510 3508	0.61638329669 38054	0.40444625214 450397	0.66873613941 9124	Normal
11	01/01/2024 05:00	0.29776371786 458966	0.74089177668 89805	0.63157058715 50474	0.42305077003 13165	0.62271372725 16767	Normal
12	01/01/2024 05:30	0.29737837440 49572	0.62483740299 07721	0.62120589478 03161	0.64919492498 97804	0.60701718506 69099	Normal
13	01/01/2024 06:00	0.41532704526 10057	0.52757262505 27567	0.30270921401 11161	0.57359783171 84226	0.58756933366 3128	Normal
14	01/01/2024 06:30	0.05611995922 370039	0.49449627186 638334	0.42760417055 13646	0.60031434352 09295	0.42285520560 424716	Normal
15	01/01/2024 07:00	0.08751369458 11612	0.53667492944 24203	0.61702904294 84833	0.56101873286 72708	0.35139775252 693606	Normal
16	01/01/2024 07:30	0.28128541179 317124	0.45771416356 31019	0.50218622594 3571	0.57616727123 55047	0.33566837019 607454	Normal
17	01/01/2024 08:00	0.20619481327 75961	0.24393231104 086152	0.33828933493 55171	0.38009820240 27808	0.21142281486 913506	Normal
18	01/01/2024 08:30	0.42737455543 254566	0.80045814013 51554	0.56343638935 00474	0.43882081589 67799	0.18969998873 41737	Normal
19	01/01/2024 09:00	0.22366265407 979818	0.54723876215 80745	0.39103180490 07188	0.34388745918 73976	0.16867879809 474662	Normal
20	01/01/2024 09:30	0.13961604977 74514	0.57982598862 74272	0.75980775474 24569	0.43739487470 31249	0.07497547454 465336	Abnormal
21	01/01/2024 10:00	0.61927479482 02592	0.51862250714 36685	0.72311383624 86436	0.54471865365 70369	0.16744224583 912914	Abnormal
22	01/01/2024 10:30	0.33737061658 557743	0.42773971835 503943	0.58486040441 46706	0.47879286986 18097	0.17894403777 203505	Normal
23	01/01/2024 11:00	0.38625470078 13206	0.50144240962 89476	0.69609715034 79358	0.61998454895 46189	0.17058281472 04016	Normal
....
50	14/04/2024	0.23365573366	0.61287204123	0.47348061864	0.26830620489	0.55528628691
00	03:30	158193	08701	331487	682694	78785	Normal

Source: (Research Results, 2025)

Data Cleansing (Data Preprocessing)

At this stage, preprocessing is performed to ensure that the dataset is ready for modeling. Several steps are involved in this process, including:

1. Missing Values

Missing values in numeric variables are handled using the Mean Imputation method, while for categorical variables, the Most Frequent Value method is used. This process is carried out using the Missing Value Node in KNIME. There are no empty values in the target variable, so there is no data row deletion process.

2. Outlier Detection and Handling

Outliers in environmental variables are identified and adjusted using Numeric Outliers. The input variables include temperature, humidity, and wind speed, while

the target and lag variables are not modified so that the time series pattern is maintained.

3. Initial Column Filtering

Several columns that are irrelevant to the prediction process, such as anomaly labels and other supporting columns, are filtered using the Column Filter Node. Only variables required for modeling are retained.

Data Transformation

This data transformation stage is carried out to prepare the dataset in a format suitable for time series analysis and regression modeling. Three main processes are carried out at this stage, including:

1. Timestamp Parsing

The Timestamp variable undergoes conversion from string format to Date&Time



format using the String to Date&Time Node. This conversion is necessary so that KNIME can recognize time data and enable temporal feature extraction in the next stage.

2. Sorting Data Based on Time (Sorting)

At this stage, the dataset is sorted chronologically using the Sorter Node to maintain the time series order. This step is important to ensure that the creation of lag features later runs correctly and avoids data leakage.

3. Data Normalization

In this study, normalization was performed by selecting the Z-score method, which was applied to numeric variables using the Normalizer Node. The goal was to balance the scale between variables so that the machine learning algorithm could learn data patterns optimally.

Feature Engineering

Feature engineering is performed to improve the model's ability to capture temporal patterns of electricity consumption. Several processes are carried out at this stage, including:

1. Creation of Lag Features

In this study, lag features were created using the Lag Column Node by generating three lags from the Electricity_Consumed variable, namely Lag-1, Lag-2, and Lag-24. Lag is the electricity consumption value at a previous time ($t-1$, $t-2$, $t-24$) which is used as a predictor for the consumption value at the current time. Lags are applied because the electricity consumption dataset is time series, where values at a given time are greatly influenced by usage patterns in the previous period. Lag-1 and Lag-2 are used to represent short-term dependency. Meanwhile, Lag-24 is used to capture daily patterns because consumption at a certain hour often resembles consumption at the same hour on the previous day. The creation of these lags will help machine learning models recognize temporal patterns and electricity consumption dynamics more accurately.

2. Temporal Feature Extraction

Hour and day of week features are extracted from the timestamp to add context to consumption behavior based on time patterns. With the Date&Time Part Extractor Node, this feature helps the model understand variations in electricity consumption based on time.

3. Feature Selection

After all new features are formed, researchers then use the Column Filter Node to remove the

original timestamp column, remove Anomaly Labels and other unused variables, and only retain features that are relevant for modeling.

Model Development

The model development stage was carried out to build an electricity consumption prediction model using three different regression algorithms. Before the training process, the dataset was divided using Partitioning Node into 80% training data and 20% test data, with random seed selection so that the experimental results could be replicated. The three machine learning algorithms used were Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression. To ensure stable model performance, the researchers also conducted a 10-fold cross-validation test using the X-Partitioner and X-Aggregator Node. Each model was trained using the feature engineering output variables that had been previously determined.

Model Evaluation

The evaluation was conducted to assess the model's ability to predict electricity consumption. In this study, the evaluation was performed using the Numeric Scorer Node in KNIME. Three evaluation metrics were used in this study, including:

1. Root Mean Square Error (RMSE)

This RMSE value is evaluated to measure the average error in units that are the same as the target variable.

2. Mean Absolute Error (MAE)

The MAE value provides an overview of the absolute error without being influenced by extreme outliers.

3. Coefficient of Determination (R^2)

Meanwhile, R^2 shows how much of the target variability can be explained by the model.

The evaluation was conducted on the cross-validation results (to assess model generalization) and then on the test data (to assess the final performance of the model objectively). The best model was then determined based on the lowest RMSE and MAE values and the highest R^2 value.

This study focuses on regression-based models to provide a balance between predictive performance and model interpretability. Compared to more complex approaches such as deep learning models (e.g., LSTM or GRU), regression-based models require lower computational resources and are easier to interpret, making them more suitable for practical implementation in real-world scenarios. While advanced models may offer potential improvements, the results of this study demonstrate that simpler models are able to provide stable and consistent performance. In

In addition, the ensemble models used in this study were implemented using standard parameter configurations provided by the modeling framework, which are sufficient to ensure stable baseline performance. Future research may explore deep learning or hybrid approaches to further enhance prediction accuracy.

Influential Factor Analysis

At this stage, an influential factor analysis was conducted to identify which variables most influence electricity consumption levels. Two approaches were used in this study:

1. **Analysis of Linear Regression Coefficients**
Regression coefficients are analyzed to determine the direction of the variable's influence (positive/negative) and the relative contribution of the variable to electricity consumption. This approach is used because linear regression is interpretable.
2. **Feature Importance of Tree-Based Models**
Random Forest and Gradient Boosted Trees generate feature importance scores that indicate the level of contribution of each feature in the prediction process. These scores provide an overview of the model's most relevant features, especially in non-linear relationships.

The results of these two approaches will then be compared to identify the dominant factors that consistently influence electricity consumption.

RESULTS AND DISCUSSION

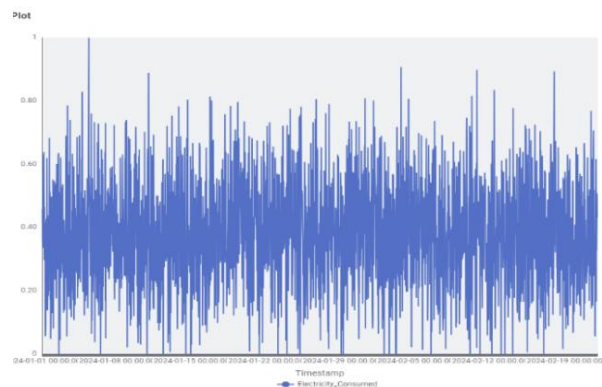
Descriptive Analysis of Datasets

The electricity consumption dataset used in this study consists of 5,000 observations with a time interval of every 30 minutes. In general, the electricity consumption pattern shows time-dependent characteristics, where the value at a given time is influenced by previous consumption. The Electricity_Consumed value varies from very low values in the early hours of the morning to peak values at certain hours, indicating a strong daily pattern. In addition to consumption variables, the dataset also includes environmental variables such as Temperature, Humidity, and Wind_Speed. Descriptively, temperature shows slight seasonal variations, while humidity and wind speed tend to fluctuate without a dominant pattern. Furthermore, to quantitatively analyze inter-variable relationships, Spearman's rank correlation analysis was conducted. The results show that environmental variables such as Temperature, Humidity, and Wind_Speed exhibit very weak correlations with electricity consumption ($|\rho| <$

0.1), indicating a limited direct relationship. In contrast, historical consumption features demonstrate stronger correlations. For instance, Avg_Past_Consumption shows a moderate positive correlation with electricity consumption ($\rho \approx 0.30$), indicating its relevance in capturing consumption patterns, indicating that past consumption contributes to predicting future electricity usage.

Further analysis using lag-based features reveals that short-term dependencies are more influential than long-term ones. Lag variables such as Electricity_Consumed(-1) and Electricity_Consumed(-2) exhibit moderate correlations ($\rho \approx 0.29$), while longer lag features such as Electricity_Consumed(-24) and Electricity_Consumed(-96) show negligible correlations ($|\rho| \approx 0.00$). This indicates that electricity consumption is primarily influenced by recent observations rather than long-term historical values. These findings are consistent with the overall correlation analysis, where historical consumption features exhibit higher correlation values with electricity consumption compared to environmental variables. In contrast, environmental variables such as Temperature, Humidity, and Wind_Speed show near-zero correlation values ($|\rho| < 0.05$), indicating a limited contribution to short-term consumption patterns.

This explains why the prediction model relies more heavily on historical consumption features, as they better capture the temporal dependencies present in the data. Based on this analysis, regression-based models are considered appropriate for this study, as they are capable of capturing linear relationships and short-term temporal dependencies present in the data. The correlation analysis also supports the feature selection strategy and provides a foundation for determining the structure of the forecasting models used in subsequent stages.

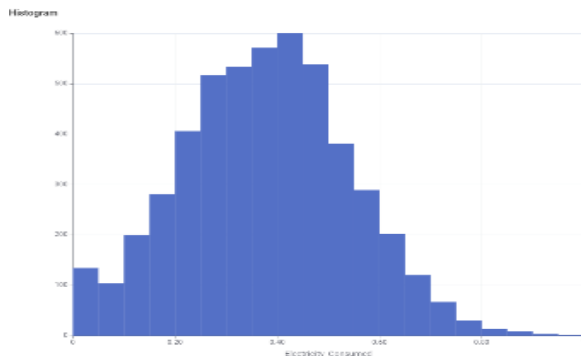


Source: (Research Results, 2025)

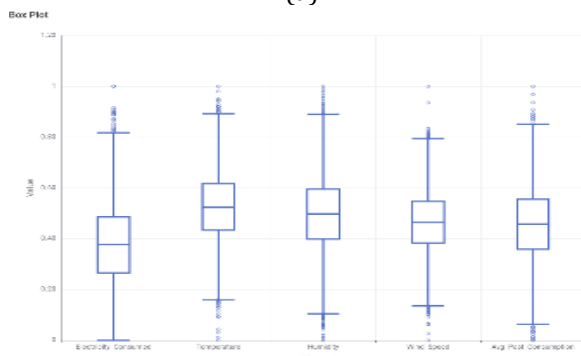
Figure 2. Electricity Consumption Patterns over Time



Figure 2 shows significant fluctuations in electricity consumption over time, reflecting the dynamic nature of time series data and its dependence on previous values. This pattern indicates a strong temporal dependence, meaning that modeling electricity consumption requires an approach capable of capturing historical information.



(a)



(b)

Source: (Research Results, 2025)
Figure 3. (a)(b) Histogram Of Electricity_Consumed Showing Data Distribution And Outliers

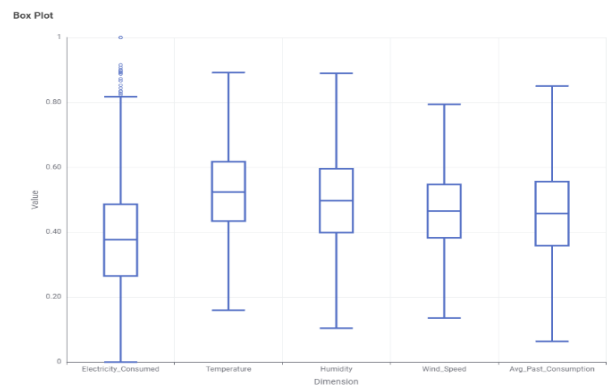
Meanwhile, Figure 3 shows that electricity consumption values are concentrated in a certain range with a distribution that is not entirely symmetrical. In addition, there are still extreme values that indicate quite high consumption variation. This condition reinforces the need for a normalization process before modeling is carried out. Furthermore, the box plot display shows that all numerical variables have a wide range of values and contain outliers. The presence of outliers in this raw data indicates variations in consumption conditions and environmental factors, requiring outlier handling during the pre-processing stage to ensure the model works more stably. Based on the results of descriptive analysis and visualization of the raw data, it can be concluded that the dataset has complex time series characteristics with high value variations, requiring a modeling approach that can

capture temporal patterns while maintaining model stability.

Data Pre-processing Results and Feature Engineering.

The data pre-processing stage is carried out to improve the quality of the dataset before it is used in the modeling process. This stage includes handling missing values, handling outliers, data transformation, and normalization of numerical variables. This process is important to ensure that the data is in a stable condition and in accordance with the assumptions of the machine learning algorithm used. Preliminary examination results indicate that most numerical variables do not have significant missing values. Missing values were handled using the mean imputation method for numerical variables and the most frequent value method for categorical variables. This approach was chosen because it preserves the characteristics of the data distribution without significantly reducing the number of observations.

Next, outlier handling was performed on environmental variables, namely Temperature, Humidity, and Wind_Speed, using the Numeric Outliers method. The target variable (Electricity_Consumed) was not included in this process to maintain the authenticity of the time-dependent electricity consumption pattern. The following is a visualization of the data after the outlier handling process was performed.



Source: (Research Results, 2025)
Figure 4. Data Visualization After Outlier Handling

Figure 4 shows the results of the box plot of numerical variables after the outlier handling process. It can be seen that the environmental and historical consumption variables have a more stable distribution with fewer extreme values. Meanwhile, the Electricity_Consumed variable still shows the presence of outliers. This is because the variable is a target time series, so extreme values are retained to maintain the characteristics of electricity



consumption patterns, especially during peak load conditions. Furthermore, data transformation was also carried out by converting the Timestamp column to Date&Time format and sorting the data chronologically. This step ensured that the time series structure was maintained and prevented data leakage during the modeling stage. Z-score normalization was applied using the Normalizer node in KNIME, where each numerical feature was standardized based on its mean (μ) and standard deviation (σ) computed from the dataset. The normalization parameters (μ and σ) were automatically calculated for each feature.

This transformation ensures that the normalized variables have a mean close to zero and a standard deviation of one. The normalization process was applied only to numerical input variables, including Temperature, Humidity, Wind_Speed, Avg_Past_Consumption, and lag features of electricity consumption. The target variable (Electricity_Consumed) and temporal features (Hour and Day of Week) were excluded from normalization to preserve their original interpretability and prevent distortion of the prediction target. During the feature engineering stage, Lag feature selection was based on the temporal resolution and characteristics of the dataset. Since the data is recorded at 30-minute intervals, lag-1 and lag-2 were selected to capture immediate short-term dependencies between consecutive observations. Lag-24 was chosen to represent consumption behavior within a broader temporal window, providing additional historical context while maintaining a manageable feature set.

The selection of these lag values was guided by the need to balance model complexity and the ability to capture relevant temporal patterns, avoiding excessive lag features that may introduce noise or overfitting. These preprocessing and feature engineering steps contribute to improving model performance by enhancing data consistency and enabling the model to capture temporal dependencies more effectively. Normalization ensures stable learning across features, while lag

features provide essential historical information that supports the model in learning electricity consumption patterns more accurately. In addition, temporal features such as hour and day of week successfully added temporal context to electricity consumption behavior.

These features enabled the model to recognize differences in consumption patterns based on the time of day and day of the week. Once all features were formed, feature selection was performed to retain only variables relevant to the modeling process, thereby controlling model complexity without sacrificing important information. Overall, the pre-processing and feature engineering results show that the combination of historical and temporal features produces a more informative data representation.

This stage is an important foundation for the development of an electricity consumption prediction model and contributes directly to improving model performance in the subsequent evaluation stage. In this study, the modeling approach focuses on comparing multiple regression-based methods under the same experimental setting rather than including baseline models such as naive or seasonal persistence approaches. This allows a consistent evaluation of model performance across different algorithms. Furthermore, while data preprocessing has addressed missing values and outliers, this study does not explicitly explore data imbalance characteristics.

The dataset used represents a single region and a specific time period; therefore, the results may be influenced by the inherent characteristics of the data and have not been validated on diverse datasets. Although temporal features such as hour and day of week are included to capture recurring patterns, broader seasonality effects and unexpected external shocks are not explicitly modeled in this study. Future work may consider incorporating baseline comparisons, additional contextual variables, and more diverse datasets to enhance model robustness and generalization.

Table 3. Statistics of Preprocessing and Feature Engineering Results Dataset

Name	Type	Missing Value	Unique Values	Minimum	Maximum	25% Quantile	50% Quantile	75% Quantile	mean	mean absolute value	Standard Deviation	Sum
Electricity_Consumed	Number (Float)	0	4945	0.0	1.0	0.265	0.377	0.486	0.377	0.131	0.164	1.883
Temperature	Number (Float)	0	4968	2.714	2.723	-0.675	-0.008	0.684	-0	0.799	1	-0
Humidity	Number (Float)	0	4955	2.703	2.695	-0.678	-0.001	0.671	0	0.798	1	0



Name	Type	Missing Value	Unique Values	Minimum	Maximum	25% Quantile	50% Quantile	75% Quantile	mean	mean absolute value	Standard Deviation	Sum
Wind_Speed	Number (Float)	0	4966	2.749	2.757	-0.684	-0.002	0.692	0	0.799	1	0
Avg_Past_Consumption	Number (Float)	0	4975	2.704	2.684	-0.683	-0.006	0.663	-0	0.797	1	-0
Electricity_Consumed(-1)	Number (Float)	1	4944	2.295	3.797	-0.678	0.004	0.666	0	0.802	1	0
Electricity_Consumed(-2)	Number (Float)	2	4943	2.295	3.797	-0.678	0.003	0.666	0	0.802	1	0
Electricity_Consumed(-24)	Number (Float)	24	4921	2.297	3.803	-0.679	0.004	0.667	0	0.802	1	0
Hour	Number (Integer)	0	24	0.0	23.0	5.0	11.0	17.0	11.48	6.006	6.929	57.42
Day of Week (Number)	Number (Integer)	0	7	1.0	7.0	2.0	4.0	6.0	4.24	1.707	1.99	20.12

Source: (Research Results, 2025)

Table 3 shows a summary of descriptive statistics from the final dataset used in the modeling stage. All numerical variables have undergone pre-processing, including missing value handling, Z-score normalization, and temporal and lag feature engineering. The minimum and maximum values for most variables are within a relatively balanced negative and positive range, indicating that the normalization process has successfully equalized the scale between features.

The Electricity_Consumed variable as the prediction target has an average value of 0.377 with moderate variation, reflecting the dynamics of electricity consumption that fluctuates over time. Environmental variables such as Temperature, Humidity, and Wind_Speed have mean values close to zero and standard deviations close to one, which is consistent with the characteristics of Z-score normalized data. The historical feature Avg_Past_Consumption and the lag features (Electricity_Consumed Lag-1, Lag-2, and Lag-24) show a relatively balanced distribution, indicating that past consumption information has been successfully retained as a potential predictor in the regression model. Although a small number of missing values are found in the lag variables, this condition is a natural consequence of the lag creation process in time series data.

These missing values appear in the initial observations due to the unavailability of historical data at earlier time steps (for example, at the beginning of the series for Lag-24). In this study, missing values in the original dataset were previously handled using mean imputation for numerical variables and the most frequent value for

categorical variables. However, missing values generated from the lag construction process were handled by excluding the affected records from the modeling stage. This approach was chosen to maintain data integrity and avoid introducing artificial bias through imputation, considering that the number of affected observations is relatively small and does not significantly impact model performance. In addition, the Hour and Day of Week temporal features show an even distribution, which allows the model to capture variations in electricity consumption based on daily and weekly time patterns. Overall, these descriptive statistics indicate that the final dataset is ready and stable for use in the development and evaluation stages of the prediction model.

Modeling Results and Model Evaluation

At this stage, electricity consumption prediction modeling was carried out using three machine learning regression algorithms, namely Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression. All models were trained using a pre-processed and feature-engineered dataset, which included historical features (lag features), temporal features, and environmental variables. The dataset was divided into 80% training data and 20% test data, and 10-fold cross-validation was performed to assess the stability and generalization ability of the model. Model performance was evaluated using three metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

Table 4. Results of the Performance Evaluation of the Electricity Consumption Prediction Model

Evaluation Matrix	Linear Regression		Random Forest Regression		Gradient Boosted Trees Regression	
	Training	Testing	Training	Testing	Training	Testing
R ²	0.121	0.140	0.091	0.09	0.087	0.105
mean absolute error	0.123	0.125	0.124	0.129	0.125	0.128
mean squared error	0.023	0.024	0.024	0.026	0.024	0.025
root mean squared error	0.153	0.156	0.155	0.161	0.156	0.159
mean signed difference	-0.001	0.001	0.000	0.001	-0.001	0.001
adjusted R ²	0.121	0.140	0.092	0.09	0.087	0.105

Source: (Research Results, 2025)

Table 4 presents a comparison of the performance of three regression models, namely Linear Regression, Random Forest Regression, and Gradient Boosted Trees Regression, based on training (cross-validation) and testing results. The evaluation was carried out using several numerical metrics to assess the error rate and the model's ability to explain the variance in electricity consumption data.

The performance of the electricity consumption prediction model was evaluated using several numerical metrics, namely Coefficient of Determination (R²), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Signed Difference (MSD). The evaluation in this study focused on the results of R², MAE, and RMSE. Furthermore, the evaluation was carried out in two scenarios, namely the training results using cross-validation and the testing results on the testing data, to assess the generalization ability of each model.

Table 5. Comparison of R² Results

Model	R ² Training	R ² Testing
Linear Regression	0.121	0.140
Random Forest	0.091	0.090
Gradient Boosted Trees	0.087	0.105

Source: (Research Results, 2025)

Based on the evaluation results shown in Table 5, the R² values obtained by the three models are relatively low, with a maximum value of 0.140 in Linear Regression. These low R² values indicate that only a small portion of the variation in electricity consumption can be explained by the model.

However, due to the dynamic and behavior-driven nature of electricity usage, not all influencing factors are fully represented in the dataset. This condition is a characteristic commonly found in household electricity consumption data, where energy usage patterns are greatly influenced by human behavior, which is dynamic and difficult to model deterministically.

To provide a more practical interpretation, error-based metrics such as RMSE and MAE are emphasized as primary evaluation measures, as they directly reflect the magnitude of prediction errors. Previous studies in time-series forecasting have highlighted that RMSE and MAE are more suitable for evaluating real-world predictive performance compared to R², particularly in datasets with high variability and stochastic behavior [22] [23]. When compared to the scale of electricity consumption values, the obtained RMSE values indicate that the prediction errors remain relatively small and consistent across models.

In addition, there are many latent factors that are not recorded in the dataset, such as occupant activities, the use of certain electrical appliances, and daily habits, which also contribute to the low R² value. However, Linear Regression showed the highest R² value in the testing data compared to Random Forest Regression and Gradient Boosted Trees Regression. In addition, the difference in R² values between the training data and the testing data was relatively small, indicating that the model did not experience overfitting and was able to generalize quite well.

This shows that a simple linear model is able to capture global patterns of electricity consumption in a stable manner. These results indicate that, despite the inherent variability in electricity consumption data, the models are able to provide stable and consistent predictions. This is supported by the relatively small differences in RMSE values across models, as well as the results of the paired t-test, which show no statistically significant differences in prediction errors. This level of performance is considered sufficient for capturing general consumption patterns and supporting short-term electricity forecasting.

Table 6. Comparison of Mean Absolute Error (MAE) Results

Model	MAE Training	MAE Testing
Linear Regression	0.123	0.125
Random Forest	0.124	0.129
Gradient Boosted Trees	0.125	0.128

Source: (Research Results, 2025)

Based on Table 6, in terms of prediction error, Linear Regression also produced the lowest



Mean Absolute Error (MAE) value in the testing data. MAE measures the average absolute error between the predicted value and the actual value, thus providing a direct picture of how far the prediction deviates from the actual value. The lower MAE value in Linear Regression indicates that this model produces more consistent predictions on the average electricity consumption scale compared to the other two ensemble models.

Table 7. Comparison of Root Mean Squared Error (RMSE) Results

Model	RMSE Training	RMSE Testing
Linear Regression	0.153	0.156
Random Forest	0.155	0.161
Gradient Boosted Trees	0.156	0.159

Source: (Research Results, 2025)

Table 7 shows the evaluation results using RMSE, indicating that the differences in performance between models are relatively small. However, to ensure that these differences are statistically valid, a paired t-test was conducted to compare the prediction errors between models. This analysis is different from the t-test used in regression coefficients, which assesses the significance of individual variables. The results show that all pairwise comparisons—between Linear Regression and Random Forest ($t = -1.7145$, $p\text{-value} = 0.0865$), between Linear Regression and Gradient Boosted Trees ($t = -1.2271$, $p\text{-value} = 0.2199$), and between Random Forest and Gradient Boosted Trees ($t = 0.4663$, $p\text{-value} = 0.641$)—produce p-values greater than 0.05.

This indicates that there are no statistically significant differences in prediction performance among the evaluated models. These findings suggest that all models demonstrate consistent and stable predictive behavior in capturing general electricity consumption patterns. Similar observations have been reported in previous studies [24], where comparable error values across models indicate robustness in handling time-series data, even under high variability conditions. In addition to RMSE, MAE and R^2 , the Mean Signed Difference (MSD) metric is used to evaluate the average bias of the predictions. MSD is calculated as:

$$MSD = (1/n) \sum (\hat{y} - y) \quad (1)$$

Where \hat{y} represents the predicted value and y represents the actual value. This metric indicates whether a model tends to systematically overestimate (positive MSD) or underestimate (negative MSD) the target variable. The MSD results for all models are close to zero, indicating that there is no systematic bias in the predictions. This

suggests that the models produce balanced predictions without consistently overestimating or underestimating electricity consumption. Overall, although slight differences in RMSE values are observed, the statistical analysis confirms that these differences are not significant. Therefore, all models demonstrate comparable performance and are equally capable of capturing electricity consumption patterns. Under these conditions, simpler models such as Linear Regression provide practical advantages due to their interpretability and computational efficiency.

Table 8. Analysis Results of Influential Factor Coefficients

Variable	Coeff.	Std. Err.	t-value	P> t
Temperature	-0.003	0.0027	-11.147	0.2651
Humidity	0.0023	0.0027	0.8696	0.3846
Wind_Speed	-0.0002	0.0027	-0.0767	0.9388
Avg_Past_Consumption	0.061	0.003	200.835	0.0
Electricity_Consumed(-1)	-0.0237	0.0029	-82.332	2.22
Electricity_Consumed(-2)	-0.0191	0.0029	-6.536	7.34
Electricity_Consumed(-24)	0.0018	0.0027	0.6593	0.5098
Intercept	0.3779	0.0027	1.395.214	0.0
R-Squared:	0.1144			
Adjusted R-Squared:	0.1125			

Source: (Research Results, 2025)

Based on Table 8, the results of the Linear Regression coefficient analysis show that historical consumption-based variables have the most significant effect on electricity consumption. This is indicated by the coefficient value and statistical significance (p-value) in the lag feature and historical consumption average. The Avg_Past_Consumption variable has the largest positive coefficient and is significant ($p\text{-value} \approx 0$), indicating that an increase in average historical consumption directly increases electricity consumption in the current period.

This finding indicates that past energy usage patterns are a strong indicator for predicting future consumption. The Electricity_Consumed(-1) and Electricity_Consumed(-2) features also show a significant effect with negative coefficients and very small p-values. This indicates the existence of short-term adjustment dynamics, where spikes in consumption in the previous period tend to be followed by a relative decline in the subsequent period. Meanwhile, Electricity_Consumed(-24)



does not show a statistically significant effect, indicating that daily patterns are not always consistent for each time interval. Conversely, environmental variables such as Temperature, Humidity, and Wind_Speed have small coefficients with relatively large p-values, so their contribution to electricity consumption predictions is limited.

Table 9. Feature Importance Analysis Results

Variabel	Split s(le vel 0)	Split s(le vel 1)	Split s(le vel 2)	Candi dates(level 0)	Candi dates(level 1)	Candi dates(level 2)
Temper ature	3	8	23	29	56	109
Humidit y	5	17	32	38	82	134
Wind_Sp eed	8	14	35	31	63	133
Avg_Pas t_Consum ption	38	50	65	41	70	110
Electrici ty_Consum ed(- 1)	11	26	48	27	67	111
Electrici ty_Consum ed(- 2)	10	20	35	35	58	106
Electrici ty_Consum ed(- 24)	24	37	55	37	74	126
Hour Day of Week (Numbe r)	1	1	12	33	55	120
	0	2	8	29	75	101

Source: (Research Results, 2025)

Based on the results of Table 9, factor analysis was also performed using feature importance from Random Forest Regression. Based on Attribute Statistics, historical consumption-based features such as Avg_Past_Consumption and electricity consumption lag (lag-1, lag-2, and lag-24) showed the greatest contribution to the prediction process. These features are most often used as separating nodes in the tree structure, indicating the dominant role of historical information in determining current electricity consumption. Conversely, environmental variables such as temperature, humidity, and wind speed have a relatively smaller contribution. These results are consistent with the Linear Regression coefficient analysis and confirm that historical consumption patterns are a major factor in short-term electricity consumption prediction.

Overall, the results of the influential factor analysis show that electricity consumption in the previous period is the main determinant in

predicting short-term electricity consumption. The consistency of findings between the Linear Regression and Random Forest Regression models reinforces the conclusion that historical information plays a more dominant role than environmental variables. These findings confirm that time series-based electricity consumption modeling needs to emphasize historical features as the main component, while external variables serve as supporting factors in improving prediction accuracy.

CONCLUSION

This study demonstrates that regression-based machine learning approaches can effectively predict short-term electricity consumption using time series smart meter data. The experimental results show that all evaluated models achieve relatively similar performance, with RMSE values in the testing phase ranging from 0.156 to 0.161. These results indicate that each model is capable of capturing electricity consumption patterns with comparable accuracy. Further analysis reveals that historical consumption features play a dominant role in improving prediction performance, while environmental variables contribute less significantly. In addition, statistical testing using the paired t-test confirms that the differences in prediction errors among the models are not statistically significant.

These findings suggest that increasing model complexity does not necessarily provide substantial performance improvement for this type of data. Therefore, model selection should consider not only predictive performance but also interpretability and computational efficiency. In this context, Linear Regression offers practical advantages while maintaining competitive accuracy. The main contribution of this study lies in demonstrating the effectiveness of simple regression-based models combined with time-series feature engineering for electricity consumption prediction, as well as providing empirical evidence that model performance differences are statistically insignificant. Future research may explore the integration of additional contextual variables and advanced modeling approaches to further enhance prediction accuracy and generalization.

REFERENCE

- [1] T. Dev, M. Haghiri, and G. Sabau, "Impacts of Urbanization on Energy Consumption in the South Asian Association for Regional



- Cooperation Zone," *Sustain.*, vol. 16, no. 18, 2024, doi: 10.3390/su16188141.
- [2] A. A. Pierre, S. A. Akim, A. K. Semenyó, and B. Babiga, "Peak Electrical Energy Consumption Prediction by ARIMA, LSTM, GRU, ARIMA-LSTM and ARIMA-GRU Approaches," *Energies*, vol. 16, no. 12, 2023, doi: 10.3390/en16124739.
- [3] K. Misiurek, T. Olkuski, and J. Zy'sk, "Review of Methods and Models for Forecasting Electricity Consumption," *Energies*, vol. 18, no. 15, p. 4032, 2025, doi: 10.3390/en18154032.
- [4] J. Yang, M. Tuo, J. Lu, and X. Li, "Analysis of Weather and Time Features in Machine Learning-aided ERCOT Load Forecasting," *2024 IEEE Texas Power Energy Conf. TPEC 2024*, 2024, doi: 10.1109/TPEC60005.2024.10472183.
- [5] A. G. Koumentakos, "Electricity Consumption Timeseries Forecasting: Statistical, Machine- and Deep-Learning Methods," *M.Sc. thesis, Dept. Informatics and Statistics, Athens Univ. of Economics and Business (AUEB), Athens, Greece*, Nov. 2022.
- [6] A. T. Mustafa and O. S. A. D. Al-Yozbaky, "Forecasting energy demand and generation using time series models: A comparative analysis of classical, grey, fuzzy, and intelligent approaches," *Franklin Open*, vol. 12, no. August, p. 100350, 2025, doi: 10.1016/j.fraope.2025.100350.
- [7] E. Demir and S. Gunal, "Short-term electricity consumption forecasting with deep learning," *J. Supercomput.*, vol. 81, no. 10, pp. 1-30, 2025, doi: 10.1007/s11227-025-07564-5.
- [8] C. Ragupathi, S. Dhanasekaran, N. Vijayalakshmi, and A. O. Salau, "Prediction of electricity consumption using an innovative deep energy predictor model for enhanced accuracy and efficiency," *Energy Reports*, vol. 12, no. September, pp. 5320-5337, 2024, doi: 10.1016/j.egy.2024.11.018.
- [9] 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," *Proc. - RoEduNet IEEE Int. Conf.*, vol. 2021-November, 2021, doi: 10.1109/RoEduNet54112.2021.9638276.
- [10] R. Mohan and P. Nikhil, "An ensemble model for the energy consumption prediction of residential buildings," *Energy*, vol. 314, p. 134255, 2025, doi: https://doi.org/10.1016/j.energy.2024.134255.
- [11] Y. Chen, W. Gong, C. Obrecht, and F. Kuznik, "A review of machine learning techniques for building electrical energy consumption prediction," *Energy AI*, vol. 21, no. April, p. 100518, 2025, doi: 10.1016/j.egyai.2025.100518.
- [12] A. Diamah, E. Sandi, Soeprijanto, S. Kusumawardhani, and M. Wagner, "Electricity Consumption Forecasting in Indonesia: Methods and Factors," *2023 10th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2023*, pp. 265-270, 2023, doi: 10.1109/ICITACEE58587.2023.10277486.
- [13] Akhmad Faeda Insani, Ahmad Mushawir, Zainuddin, Aditya Adiaksa, and Sparisoma Viridi, "Optimization of Variable Combinations for Household Electricity Consumption Prediction Using a Multivariate Time Series Machine Learning Approach," *INOVTEK Polbeng - Seri Inform.*, vol. 10, no. 1, pp. 87-98, 2025, doi: 10.35314/hd6bv378.
- [14] M. Hasan *et al.*, "A state-of-the-art comparative review of load forecasting methods: Characteristics, perspectives, and applications," *Energy Convers. Manag. X*, vol. 26, no. January, p. 100922, 2025, doi: 10.1016/j.ecmx.2025.100922.
- [15] V. Arabzadeh and R. Frank, "A Four-Dimensional Analysis of Explainable AI in Energy Forecasting: A Domain-Specific Systematic Review," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 4, p. 153, 2025, doi: 10.3390/make7040153.
- [16] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," *Adv. Appl. Energy*, vol. 9, no. October 2022, p. 100123, 2023, doi: 10.1016/j.adapen.2023.100123.
- [17] M. H. Saleem *et al.*, "A comprehensive review of machine learning and deep learning models for non-intrusive load monitoring: performance, analyses, practical insights, and emerging trends," *Appl. Intell.*, vol. 55, no. 15, 2025, doi: 10.1007/s10489-025-06921-4.
- [18] Y. S. Kim, M. K. Kim, N. Fu, J. Liu, J. Wang, and J. Srebric, "Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models," *Sustain. Cities Soc.*, vol. 118, no. October 2023, p. 105570, 2025, doi: 10.1016/j.scs.2024.105570.

- [19] L. Cascone, S. Sadiq, S. Ullah, S. Mirjalili, H. U. R. Siddiqui, and M. Umer, "Predicting Household Electric Power Consumption Using Multi-step Time Series with Convolutional LSTM," *Big Data Res.*, vol. 31, p. 100360, 2023, doi: <https://doi.org/10.1016/j.bdr.2022.100360>.
- [20] J. F. Torres, F. Martínez-Álvarez, and A. Troncoso, "A deep LSTM network for the Spanish electricity consumption forecasting," *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10533–10545, 2022, doi: [10.1007/s00521-021-06773-2](https://doi.org/10.1007/s00521-021-06773-2).
- [21] T. Čegovnik, A. Dobrovoljc, J. Povh, M. Rogar, and P. Tomšič, "Electricity consumption prediction using artificial intelligence," *Cent. Eur. J. Oper. Res.*, vol. 31, pp. 833–851, 2023, doi: <https://doi.org/10.1007/s10100-023-00844-6>.
- [22] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3>
- [23] K. Misiurek, T. Olkusi, and J. Zyśk, "Review of Methods and Models for Forecasting Electricity Consumption," *Energies*, vol. 18, no. 15, Art. no. 4032, 2025, doi: [10.3390/en18154032](https://doi.org/10.3390/en18154032).
- [24] K. Maciejowska, A. Lipiecki, and B. Uniejewski, "Statistical and economic evaluation of forecasts in electricity markets: Beyond RMSE and MAE," *arXiv preprint arXiv:2511.13616*, 2025, doi: [10.48550/arXiv.2511.13616](https://doi.org/10.48550/arXiv.2511.13616)

