AUT Journal of Electrical Engineering
10.22060/EEJ.2026.25648.5974

Interpretable Hourly Electricity Demand Forecasting: Lagged Load-Temperature Effects and Model Explainability Study

Mehdi Jahangiri [1], Peyman Babaei [2, *]

[1] Energy and Environment Research Center, ShK.C., Islamic Azad University, Sharekord, Iran.
[2] Computer Department, WT.C., Islamic Azad University, Tehran, Iran.
[*] Corresponding Author: peyman.babaei@iau.ac.ir

Abstract:

Short-term electricity consumption forecasting plays a crucial role in the optimal operation of energy systems and smart grids. Despite the widespread success of machine learning models in improving forecasting accuracy, their lack of interpretability often hinders trust and application in energy decision-making. This paper presents an interpretable framework for hourly electricity consumption prediction using lagged load and temperature features. To this end, the performance of linear and non-linear models is systematically compared on a one-year hourly electricity consumption dataset, including ambient temperature and workday information for a commercial building. To enhance model transparency, Explainable Artificial Intelligence methods are employed to analyze feature importance at both global and local levels. Results indicate that short-term and daily lags of electricity consumption play a dominant role in consumption dynamics, while the effect of temperature is non-linear and often appears with a time delay. Tree-based models significantly outperform the linear model and demonstrate a higher ability to identify complex interactions and lagged effects of temperature. SHapley Additive exPlanations analysis reveals consistent patterns in feature importance, and its waterfall plots and Local Interpretable Model Agnostic Explanations provide complementary local insights under different consumption conditions. The proposed framework, while improving forecasting accuracy, provides a transparent and meaningful interpretation of electricity consumption behavior and can be used as a reliable tool in energy planning and demand-side management.

Keywords: Ridge Regression; Random Forest; XGBoost; SHAP; LIME.

The authors would like to acknowledge the use of the ChatGPT-5.5 (OpenAI) large language model for linguistic refinement and grammatical editing of the manuscript during the revision process. It is important to note that the authors reviewed and edited the AI-generated suggestions and remain fully responsible for the scientific content and final version of the paper.

10.22060/EEJ.2026.25648.5974

1. Introduction

Electricity consumption forecasting is one of the most critical issues in power system engineering and energy management, playing a pivotal role in optimal resource allocation, reducing operational costs, and preventing grid instability [1]. The ever-increasing energy consumption and fluctuations arising from the integration of renewable energy sources, such as solar and wind power, have made precise grid management more complex. Short-term load forecasting, particularly at the hourly scale, is of particular importance as daily decision-making and power generation scheduling depend on it [2, 3]. Errors in short-term forecasts can lead to over-generation or energy shortages, both of which have significant economic and operational consequences.

Initially, conventional statistical models such as ARIMA, Moving Average, and Linear models were developed for electricity time series forecasting [3]. These models are capable of identifying linear consumption and simple seasonal patterns, but due to the non-linear and complex nature of electricity consumption and its dependence on environmental and human factors, they have limitations in many real-world applications [4]. Studies show that these models do not provide sufficient accuracy when facing multi-seasonal patterns and temperature lag effects and often require strict assumptions about error distribution and data stationarity [4, 5].

One of the most important characteristics of electricity consumption time series is its dependence on past values (lagged features). Short-term lags (1 to several hours ago) and daily lags (the same hour of the previous day) play a very important role in consumption forecasting [5]. This short-term dependency indicates the lasting effect of previous consumption and user behavior patterns and is one of the most important factors in modeling network behavior [6]. In addition, environmental variables such as air temperature, humidity, and economic indicators also have significant non-linear and delayed effects on electricity consumption. For example, temperature usually affects electricity consumption with a delay of 1 to 3 hours, which is due to the thermal inertia of buildings and heating and cooling systems [7].

With advances in computing and increasing volumes of electricity consumption data, machine learning models have emerged as efficient alternatives to traditional models. Models such as Artificial Neural Networks (ANNs) have been able to improve forecasting accuracy by identifying complex non-linear

10.22060/EEJ.2026.25648.5974

patterns [8, 9]. Furthermore, tree-based decision models such as Random Forest and XGBoost, with their ability to model complex interactions between features and without requiring the assumption of data stationarity, have gained a special place [10, 11]. These models can simultaneously identify short-term lags, temperature effects, and complex interactions between features, which have shown higher efficiency than linear models in many studies.

Despite the high accuracy of non-linear models, one of their main limitations is the lack of interpretability [12, 13]. In other words, the decisions made by these models are not easily understood by engineers and network decision-makers, and this is a barrier to trusting the model and its practical application. Recent studies have shown that the lack of model transparency can reduce model acceptance and application in network management, especially in environments where economic and operational decisions require scientific explanation [7, 14].

However, although the focus of early research was on improving forecasting accuracy, in recent years, another aspect has been prominently considered, namely explainability. In fact, many machine learning and deep learning models, despite their high accuracy, are known as “black boxes” because the way they make decisions is not transparent to end-users such as network operators or energy planners. In applications such as power grids, where decisions based on model results can affect national stability and energy security, the lack of acceptable explainability is unacceptable [15].

In response to this need, the field of Explainable Artificial Intelligence (XAI) has grown rapidly in recent years, and various methods have been introduced for extracting feature importance, analyzing variable contributions, and interpreting model behavior [16]. The SHAP method, based on Shapley value theory, is known as one of the most effective methods for interpreting complex models, which can calculate the contribution of each feature to the model’s decision in a quantitative and fair manner [17]. In addition to providing overall feature importance, SHAP (SHapley Additive exPlanations) is able to provide deep insights into non-linear relationships and interaction effects and clarify model behavior at specific points [18]. This advantage distinguishes it from many previous methods, such as LIME (Local Interpretable Model Agnostic Explanations) or common Feature Importance [19].

Simultaneously with the growth of XAI, recent studies have presented combined approaches that provide both accurate prediction and reliable explainability. Also, in another paper in 2024, SHAP was used for feature engineering to increase the accuracy of the XGBoost model in predicting national energy consumption by selecting meaningful features [12]. Recent research in the field of smart buildings also shows that explainability can have a significant impact on consumption management and cost reduction. For example, a study in 2025 presented a framework based on XAI for predicting building energy consumption considering operational context (Context-Aware Forecasting). This framework, using SHAP, showed that different operational conditions, such as the presence of people, weather patterns, and changes in building technology, can provide different prediction models [20].

In this regard, XGBoost has been introduced as one of the most accurate and stable algorithms for electric load forecasting. This model, using a gradient boosting structure, has the ability to learn complex and non-linear patterns and has achieved strong performance in many scientific and industrial applications. XGBoost's effectiveness in handling structured and tabular data has made it a widely used option for energy forecasting tasks [11]. The combination of XGBoost with explainability techniques has also attracted increasing attention in recent studies [21].

Research combining XGBoost with explainable AI techniques such as SHAP has recently been explored for energy forecasting applications. These studies mainly analyze feature importance to understand the contribution of variables in prediction models [22]. In many cases, the feature sets used in electricity forecasting models consist of basic temporal and climatic variables such as hour of the day, temperature, and day-of-week indicators [23]. Previous studies have also utilized a variety of datasets ranging from short-term experimental data to real operational measurements. Real-world electricity consumption datasets often include challenges such as noise, fluctuations, missing values, and irregular behavior, which make forecasting problems more complex and realistic [24].

As shown in Table 1, recent studies in the field of Short-Term Load Forecasting (STLF) utilizing explainable AI (XAI) methods have made significant advancements. The study by Gopal et al. [25], using an ensemble algorithm and the SHAP method, investigated feature importance on real-world data with hourly resolution; however, their feature engineering remained at an intermediate level, and no analysis of feature interactions was conducted. Mupenzi et al. [26], employing XGBoost and SHAP, provided more advanced feature engineering and partially addressed interaction analysis, yet they still relied on a single XAI method. On the other hand, Wen and Liu [27], using LightGBM and a combination of SHAP and LOFO, applied advanced feature engineering to real prosumer energy data (training period: October 2021 to March 2023; test period: April to May 2023), but feature interaction analysis was omitted in that study as well. Despite these advancements, three fundamental gaps in the existing literature can be identified: first, the dominant reliance of studies on a single XAI method (typically SHAP), which may only cover one aspect of interpretability and overlook complex local phenomena; second, the lack of systematic interaction analysis, which is essential for understanding non-linear relationships and mutual dependencies in power systems; and third, limitations in advanced feature engineering or the failure to simultaneously utilize lag, calendar, statistical, and interaction features in many studies.

Table 1. Comparative summary of the proposed framework and recent XAI-based studies in short-term load forecasting (STLF).

Study	Algorithm	Nonlinear	Feature Engineering	XAI Method	Temporal Split	Data Type	Hourly Resolution	Interaction Analysis
Gopal et al. (2025) [25]	Ensemble	✓	Moderate	SHAP	✓	Real	✓	✗
Mupenzi et al. (2026) [26]	XGBoost	✓	Advanced	SHAP	✓	Real	✓	Partial
Wen & Liu (2025) [27]	LightGBM	✓	Advanced	SHAP + LOFO	✓ (2021-10 to 2023-03 train; 2023-04 to 2023-05 test)	Real	✓	✗
This Study	XGBoost	✓	Advanced	SHAP + LIME	✓	Real	✓	✓

Aiming to bridge these gaps, the present study combines a dual XAI approach (SHAP + LIME) with advanced feature engineering and explicit analysis of feature interactions on real-world hourly data using an appropriate temporal split. This combination not only enhances interpretability at both global (SHAP) and local (LIME) levels but also facilitates a more accurate identification of energy consumption patterns and a deeper understanding of the mutual influences of factors affecting electricity load. This comprehensive approach addresses the methodological shortcomings of previous studies and paves a new path for the development of reliable and interpretable STLF models.

10.22060/EEJ.2026.25648.5974

1-1- Research Gap and Main Contributions

Despite numerous studies in the field of machine learning-based electricity load forecasting, several important issues remain insufficiently addressed in the existing literature. First, in many studies, lag features are often selected empirically without providing a clear statistical basis for their identification. Second, a large portion of the research primarily focuses on improving prediction accuracy, while providing limited insights into the physical interpretation of electricity consumption behavior and the role of influential variables. Third, model interpretability is frequently performed using only a single explainable artificial intelligence (XAI) technique, and simultaneous analysis at both global and local levels has received comparatively less attention.

To address these limitations, the present research proposes an interpretable framework for electricity consumption forecasting that integrates statistical analysis, machine learning modeling, and explainable artificial intelligence into a unified structure. The main scientific contributions of this research are summarized as follows:

- Developing a feature selection framework based on autocorrelation and mutual information analysis to identify both linear and non-linear temporal dependencies.
- Providing a comprehensive comparison between linear and non-linear models using identical feature sets in order to more accurately evaluate the role of non-linear relationships in electricity consumption forecasting.
- Investigating the lagged effects of temperature on electricity consumption and identifying complex interactions between climatic variables and consumption patterns.
- Applying SHAP and LIME simultaneously to provide interpretability at both global and local levels.
- Providing a physically meaningful interpretation of electricity consumption behavior and influential factors that can support energy planning, demand-side management, and decision-making in smart grids.

It should be noted that the novelty of this research does not lie in proposing a new machine learning algorithm. Instead, the contribution of this study is in designing an interpretable decision-support

10.22060/EEJ.2026.25648.5974

framework for electricity load forecasting. Although individual components of this framework, such as XGBoost, SHAP, and LIME, have been previously reported in the literature, existing studies rarely integrate statistical-based feature selection, linear and non-linear modeling, global and local interpretability, and physical analysis of lagged temperature effects within a single coherent framework. Therefore, the primary contribution of this work lies in the systematic integration of these components and in transforming predictive model outputs into actionable insights for energy management and decision-making.

Based on the identified research gaps, this study presents an integrated and interpretable framework for electricity demand forecasting that combines statistical analysis, machine learning modeling, and explainable artificial intelligence methods within a coherent structure. As illustrated in Fig. 1, the proposed framework consists of five sequential stages. In the first stage, raw data including hourly electricity consumption, temperature, and workday information are collected and pre-processed. Next, temporal dependencies in the dataset are analyzed using the Autocorrelation Function (ACF) and Mutual Information in order to identify effective lag features related to electricity consumption and temperature. In the subsequent stage, linear and non-linear models, including Ridge Regression, Random Forest, and XGBoost are developed and trained for electricity load forecasting. The performance of these models is then evaluated and compared using standard metrics such as MAE, RMSE, and the coefficient of determination (R^2). Finally, to enhance the transparency and reliability of the predictive models, explainable artificial intelligence techniques, including SHAP and LIME, are employed to analyze model behavior at both global and local levels. In addition to achieving high prediction accuracy, this integrated framework enables the interpretation of key factors affecting electricity consumption, the identification of lagged temperature effects, and the extraction of practical insights for energy planning, demand-side management, and intelligent power system operation.

10.22060/EEJ.2026.25648.5974

Fig. 1. Workflow of the proposed interpretable hourly electricity demand forecasting framework.

2. Data Description and Preprocessing

2-1- Dataset Description and Quality Assessment

To evaluate the reliability of the dataset, a comprehensive data quality analysis was conducted prior to the modeling stage. The results showed that approximately 1.8% of the total observations contained missing values, which were mainly caused by short interruptions in sensor data recording. These values were estimated using linear time-based interpolation. In total, only 2.1% of the data underwent the imputation process, indicating that the level of preprocessing intervention was minimal.

In addition, using statistical criteria including the interquartile range (IQR) and the standard deviation threshold ($|z|>3$), approximately 4.3% of the observations were identified as outliers. Considering the nature of energy consumption data and the presence of real peaks during peak load hours, these observations were not removed entirely. Instead, a mild Winsorization approach at the 1st and 99th percentile was applied in order to control unrealistic extreme values while preserving the dynamic structure of electricity consumption.

Further inspection indicated no evidence of long-term sensor failures, gradual sensor drift, or systematic anomalies in the recorded measurements. Therefore, the dataset can be considered reliable in terms of temporal stability and measurement accuracy. Moreover, the use of tree-based models such as Random

10.22060/EEJ.2026.25648.5974

Forest and XGBoost, which are inherently robust to noise and outliers, further improves the robustness of the forecasting results.

To further examine the influence of imputed data on model performance, a sensitivity analysis was conducted in which the XGBoost model was evaluated under two conditions: using the entire preprocessed dataset (including the 2.1% imputed observations) and removing all records containing imputed values. The results showed that the R^2 value was 0.9781 when imputed data were included and 0.9764 when they were removed. The observed difference ($\Delta R^2 = 0.0017$, less than 0.2%) is statistically negligible and indicates that the high performance of the models is not dependent on the data imputation process. Overall, the combination of satisfactory data quality and the robustness of the employed models strengthens the statistical validity and reliability of the forecasting results presented in this study.

2-2- Data Preprocessing

The data used in this study consists of hourly time series of electricity consumption and ambient temperature over a one-year period. Fig. 2 shows the hourly changes in electricity consumption and ambient temperature during the study period. As can be seen, electricity consumption has regular fluctuations at different time scales, indicating the presence of repeating daily and weekly patterns. In contrast, temperature has a smoother behavior, and its changes are accompanied by a time delay relative to electricity consumption. This initial observation suggests that temperature can act as an exogenous variable with a delayed effect in predicting electricity consumption and justifies the need to examine time lags for both variables.

To ensure data quality and provide suitable conditions for training prediction models, several preprocessing steps were applied to the data in a systematic manner. In the first step, time stamps were converted to a standard date-time format and set as the time index of the data. Then, the data was synchronized to an hourly time frequency to ensure uniformity of the time interval between observations. This step is particularly important because time series-based models are sensitive to the non-uniformity of the time interval between observations.

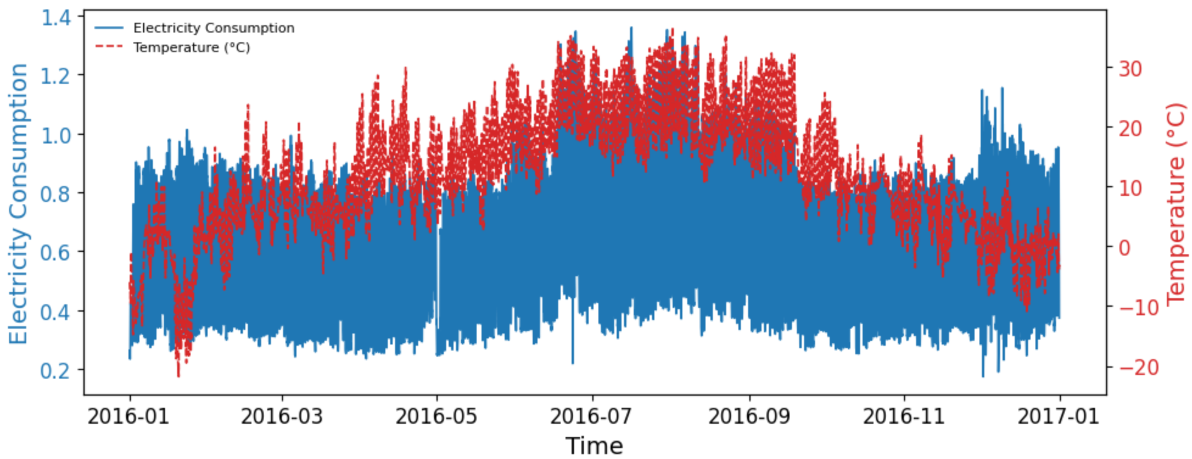


Fig. 2. Time series of hourly electricity consumption and temperature.

In order to extract the temporal dependencies, present in the electricity consumption time series and also the lagged effects of the temperature variable, a set of time-lag-based features including $Load_{t-k} = Lag_k\left(Load_t\right)$ and $Temp_{t-k} = Lag_k\left(Temp_t\right)$ were designed. The selection of these lags was based on a combination of statistical analysis of temporal dependency and examination of non-linear relationships. In the first step, considering the periodic nature of electricity consumption, an Autocorrelation Function (ACF) analysis was performed on the hourly electricity consumption time series. Fig. 3 shows the time series function of hourly electricity consumption up to a lag of 168 hours. The presence of significant autocorrelation coefficients in short-term lags, especially at lags of 1 and 2 hours, indicates a strong dependency of electricity consumption on past values in short time intervals. In addition, distinct peaks at lags of 24 and 168 hours respectively, indicate daily and weekly periodic patterns in electricity consumption. These results confirm the existence of temporal memory and periodic structure in the electricity consumption time series and provide a suitable statistical basis for selecting time lags corresponding to these periods in forecasting models.

10.22060/EEJ.2026.25648.5974

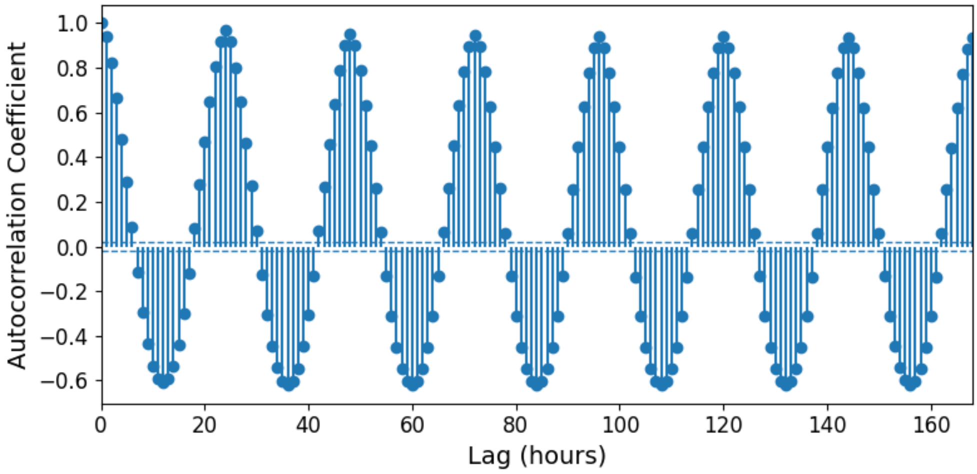


Fig. 3. Autocorrelation function of electricity consumption.

Furthermore, to confirm the obtained inference, a mutual information analysis between electricity consumption and load lags was performed. Fig. 4 shows the amount of mutual information between electricity consumption at the current time and its lagged values for different lags. Unlike ACF, which only reveals linear dependencies, mutual information has the ability to identify non-linear dependencies. The results of this figure show that the dependency between electricity consumption and its lags remains significant in short-term lags as well as in lags corresponding to daily and weekly periods. This indicates that, in addition to linear relationships, non-linear relationships also exist between current and past consumption, which can be more effectively exploited by non-linear models. Therefore, the existence of significant dependency in short-term lags and also a distinct peak in lags of 24 and 168 hours, respectively, indicating daily and weekly patterns of electricity consumption, is shown. Accordingly, the lags $Load_{t-1}$, $Load_{t-2}$, $Load_{t-24}$ and, $Load_{t-168}$ were selected as representatives of short-term and periodic dependencies of electricity consumption.

To further investigate the effect of ambient temperature on electricity consumption, a mutual information analysis was performed between electricity consumption and temperature lags. The results of this analysis showed that the dependency between temperature and electricity consumption is non-linear and lagged in nature, such that short-term lags (1 to several hours) and the lag corresponding to the daily cycle have higher values of mutual information. Accordingly, the temperature variable ($Temp_t$) and its lagged values $Temp_{t-1}$, $Temp_{t-3}$, and $Temp_{t-24}$ were selected as temperature features.

10.22060/EEJ.2026.25648.5974

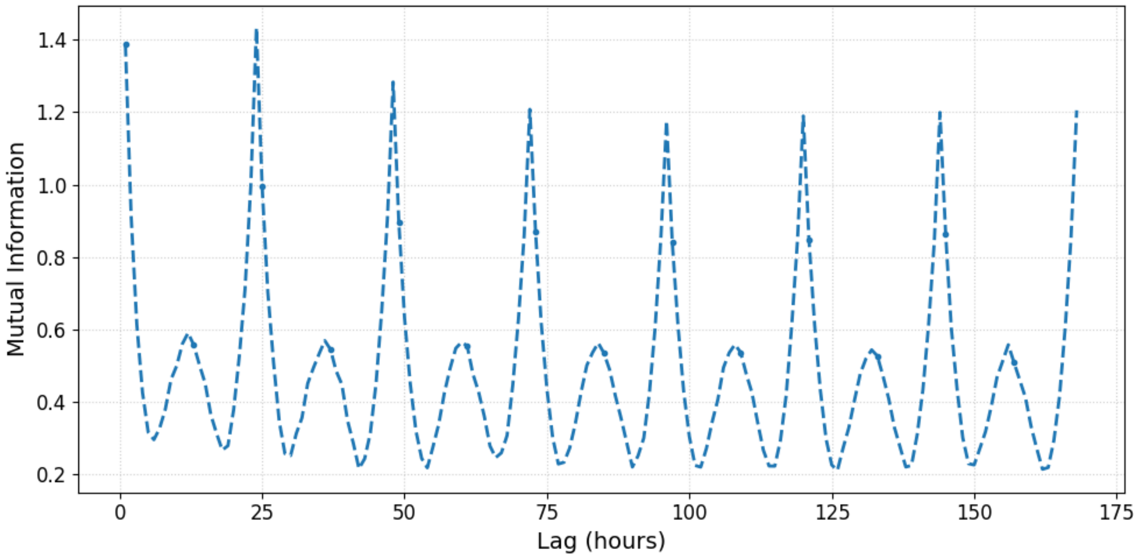


Fig. 4. Mutual information between electricity consumption and load lags.

Fig. 5 displays the mutual information between hourly electricity consumption and ambient temperature with different time lags. It is observed that the mutual information in short-term lags, especially in the range of 1 to several hours, has higher values, indicating the short-term lagged effect of temperature on electricity consumption. This behavior can be attributed to the thermal inertia of buildings and heating and cooling systems. Also, the presence of significant mutual information values at a lag of 24 hours indicates that temperature is also correlated with the daily pattern of electricity consumption. These results show that the effect of temperature on electricity consumption is non-linear and lagged, and using appropriate lags can improve the efficiency of forecasting models, especially non-linear models. To prevent excessive model complexity and reduce the risk of overfitting, only lags that were statistically and physically most justified were selected. This set of features allows for a fair comparison of linear and non-linear models while maintaining the interpretability of the results. It is worth noting that the selection of time lags in this paper was based on domain knowledge, autocorrelation analysis, and mutual information, and purely empirical or trial-and-error choices were avoided.

10.22060/EEJ.2026.25648.5974

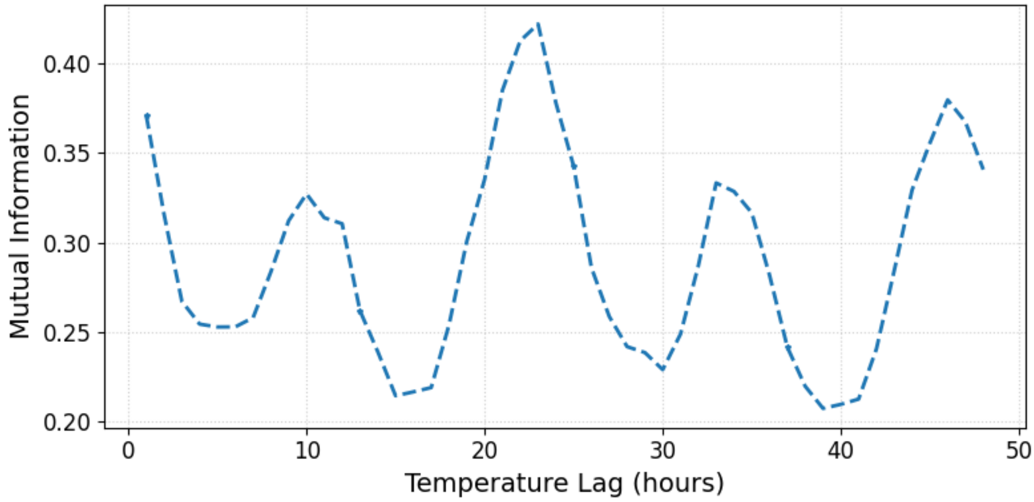


Fig. 5. Mutual information between electricity consumption and temperature lags.

In general, the autocorrelation function and mutual information analysis show that electricity consumption has a strong temporal dependency and a distinct periodic structure on daily and weekly scales, while ambient temperature, as an exogenous variable, exerts its effect on electricity consumption in a non-linear manner and with a time delay. These findings statistically and physically justify the selection of time lags for electricity consumption and temperature and provide a valid basis for designing the set of features used in linear and non-linear models. Considering workday as a binary variable, the input feature vector at time t is defined as follows, and the goal is to predict the current consumption value $y_t = Load_t$.

$$\mathbf{X}_t = \left[Load_{t-1}, Load_{t-2}, Load_{t-24}, Load_{t-168}, Temp_t, Temp_{t-1}, Temp_{t-3}, Temp_{t-24}, Workday\right] \quad (1)$$

where $\mathbf{X}_t$ represents the input feature vector at hour t. The variables $Load_{t-1}$ and $Load_{t-2}$ are included to model short-term load dependency, $Load_{t-24}$ is used to account for the daily recurring pattern, and $Load_{t-168}$ represents the weekly behavior of electricity consumption. In addition, the variables $Temp_t$, $Temp_{t-1}$, $Temp_{t-3}$, and $Temp_{t-24}$ model the direct and lagged effects of ambient temperature on energy consumption, respectively. The variable $Workday$ is also a binary variable that indicates whether the day is a working day or a non-working day. All examined models, including Ridge Regression, Random Forest, and XGBoost, were trained using this same set of input features to ensure that the comparison of model performance is based solely on their ability to learn linear and non-linear relationships and is not influenced by differences in the input variables.

10.22060/EEJ.2026.25648.5974

To model the temporal dependencies, lagged features of electricity consumption and lagged features of ambient temperature were created. These features were extracted by applying a time shift to the original time series based on the lags selected from the autocorrelation function and mutual information analysis. After creating the new dataset, the data was divided into training and testing sets. This partitioning was performed temporally and non-randomly to prevent temporal information leakage and simulate real-world forecasting conditions in practical applications.

Since linear models based on gradient optimization are sensitive to the scale of variables, the input features were first standardized, such that the mean and standard deviation were calculated only based on the training data, and then the same parameters were used to transform the test data. This approach prevents the transfer of statistical information from the test data to the training stage. It should be noted that standardization was applied only to linear models, and the preprocessing process was designed to allow for a fair comparison between linear and non-linear models while preserving the temporal structure of the data and preventing any information leakage.

3. Methodology

3-1- Ridge Regression

To evaluate the capability of linear models in capturing temporal dependencies of electricity consumption, Ridge Regression was employed as the baseline model. In Ridge Regression, the relationship between the target variable and the input features is assumed to be linear. To prevent overfitting and reduce coefficient variance, an L^2 regularization term is added to the objective function [28]. The model parameters are estimated by minimizing:

$$\sum_{i=1}^{n}\left(y_i - \mathbf{X}_i\beta\right)^2 + \lambda\beta^2 \qquad (2)$$

where λ is the regularization parameter controlling the trade-off between model complexity and goodness of fit.

3-2- Random Forest Regression

To capture non-linear relationships and complex interactions between electricity consumption and temperature features, a Random Forest regression model was developed. The same input feature set used in Ridge Regression was employed to ensure a fair comparison between linear and non-linear approaches. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and aggregates their predictions through averaging [10]:

$$\hat{f}(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x) \tag{3}$$

where B is the number of trees and $T_b(x)$ denotes the prediction of the b-th tree. At each split, the optimal partition is selected by minimizing the weighted mean squared error:

$$MSE_{split}=\frac{n_L}{n}MSE_L+\frac{n_R}{n}MSE_R \tag{4}$$

where n_L and n_R denote the number of samples in the left and right nodes, respectively.

3-3- XGBoost Model

For a more advanced modeling of non-linear interactions, the Extreme Gradient Boosting (XGBoost) algorithm was employed [11]. Similar to the previous models, identical lagged features were used as inputs. The prediction is defined as an additive ensemble of K regression trees:

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i) \tag{5}$$

The objective function consists of a loss term and a regularization term controlling model complexity:

$$L=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{6}$$

The regularization term is defined as:

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\omega^2 \tag{7}$$

where T is the number of leaves, ω represents leaf weights, and γ and λ are regularization parameters. Optimization is performed using a second-order Taylor approximation:

$$L^{(t)}\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{8}$$

where,

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i\right)}{\partial \hat{y}_i} \qquad h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i\right)^2}{\partial \hat{y}_i^2} \tag{9}$$

This formulation enables efficient learning of complex and highly nonlinear relationships.

3-4- Model Evaluation Metrics

Model performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) [29].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{12}$$

These metrics collectively evaluate prediction accuracy and generalization performance.

3-5- Statistical Tests and Residual Diagnostics

Beyond predictive accuracy, statistical diagnostics of the residuals were conducted to evaluate whether the developed models adequately captured the temporal structure of the electricity consumption data. Residual analysis is important in time-series forecasting because systematic patterns remaining in the residuals may indicate model misspecification, remaining autocorrelation, or instability in variance. Therefore, several statistical tests were applied to examine autocorrelation, distributional properties, and heteroscedasticity in the residuals.

First, the Durbin–Watson statistic was used to detect first-order autocorrelation in the residuals:

$$DW = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2} \tag{13}$$

where e_t denotes the residual at time t and n is the number of observations. Values close to 2 indicate the absence of strong first-order autocorrelation.

To examine higher-order serial dependence, the Ljung–Box test was applied:

10.22060/EEJ.2026.25648.5974

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\rho_k^2}{n-k} \tag{14}$$

where ρ_k is the autocorrelation at lag k, h is the number of lags, and n is the sample size.

The normality of residuals was assessed using the Jarque–Bera test:

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right) \tag{15}$$

where s and k represent the skewness and kurtosis of the residual distribution.

Finally, heteroscedasticity was evaluated using the Breusch–Pagan test: $BP = nR^2$, where R^2 is obtained from the auxiliary regression of squared residuals on the explanatory variables. These diagnostic tests provide additional insight into the reliability and statistical adequacy of the forecasting models.

3-6- Model Interpretability Using SHAP and LIME

To enhance model transparency, SHAP and LIME were employed for global and local interpretability. SHAP decomposes the prediction into additive feature contributions:

$$f(x) = \phi_0 + \sum_{i=1}^{M}\phi_i \tag{16}$$

The SHAP value for feature i is calculated as [13]:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(M-|S|-1)!}{M!}\left[f(S\cup\{i\}) - f(S)\right] \tag{17}$$

LIME provides local interpretability by fitting a surrogate model around a target instance:

$$\arg\min_{g\in G} L(f, g, \pi_x) + \Omega(g) \tag{18}$$

with locality weighting defined as [19]:

$$\pi_x(z) = \exp\left(-\frac{D^2(x,z)}{\sigma^2}\right) \tag{19}$$

These techniques enable analysis of feature influence at both global and instance-specific levels.

3-7- Hyperparameter Settings and Model Implementation

To ensure reproducibility, all hyperparameters are explicitly reported. The Ridge Regression regularization parameter was set to $\lambda = 1.0$. For Random Forest, 200 trees were used without restricting maximum depth, and all features were considered at each split (max_features = 1.0). For XGBoost, 500 trees were constructed with a maximum depth of 6 and a learning rate of 0.05. Subsampling ratios for observations and features were set to 0.8 to reduce overfitting. Default regularization parameters ($\lambda = 1$, $\alpha = 0$) were retained. Hyperparameters were selected using Grid Search to balance predictive accuracy and model generalization.

4. Results and Comparative Analysis

4-1- Ridge Regression model

The evaluation results of the Ridge Regression model showed that this model has good performance in both the training and testing sets, with R^2 values of 0.968 and 0.966 for the training and testing data, respectively. The proximity of these values indicates good generalization ability of the model and the effective regularization process in preventing overfitting. MAE and RMSE values were also low in both the training and testing sets, with little difference between them, indicating the stability of the model in predicting hourly electricity consumption. These results show that a significant portion of the variation in electricity consumption can be explained solely by the linear relationships between current consumption and lagged load features.

To examine the actual effect of lagged load features and determine the efficiency of the Ridge Regression model as a baseline model, a simple linear model was also developed, based on only two main features:

- Previous Hour Load: $Load_{t-1}$, this feature reflects the natural short-term dependency of electricity consumption and is considered the most effective initial linear variable for prediction in most electricity time series. Using this variable allows the impact of the continuation of previous consumption to be identified in the model.
- Current Temperature: Selected as the only external variable, reflecting the immediate environmental effect on electricity consumption. This choice provides sufficient information for a simple linear model without increasing model complexity.

10.22060/EEJ.2026.25648.5974

Evaluating the performance of the simple linear model showed that this model is capable of predicting a significant portion of electricity consumption, but its accuracy is more limited than that of the model with selected lagged load features. In the training set, MAE and RMSE were 0.0576 and 0.0870, respectively, and the R^2 was 0.8911. In the test data, MAE = 0.0565, RMSE = 0.0831, and R^2 = 0.8735. The proximity of the error values and R^2 between the training and testing sets indicates no overfitting of the model and its relatively good generalization. Comparing the simple linear model and the Ridge Regression model is presented in Table 2. The difference in R^2 between the simple linear model (0.8735) and the Ridge Regression model with selected lagged load features (0.9658) shows that adding short-term and periodic lags significantly increases the model's predictive power.

Table 2. Comparison of the simple linear model and the Ridge Regression model.

	Train			Test		
Method	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Baseline model	0.0576	0.087	0.8911	0.0565	0.0831	0.8735
Ridge Regression	0.0322	0.0472	0.968	0.0313	0.0431	0.9658

Comparing the R^2 values between the simple linear model and the Ridge Regression model with selected lags clearly demonstrates the effect of temporal lags in improving electricity consumption prediction. This analysis provides a solid scientific basis for moving to more complex models and examining the effect of non-linear features and the importance of features at both global and local levels. The evaluation results of Ridge Regression showed that the linear dependencies identified by the autocorrelation function have a dominant contribution to electricity consumption prediction, while the role of the temperature variable appears as weaker and potentially non-linear effects.

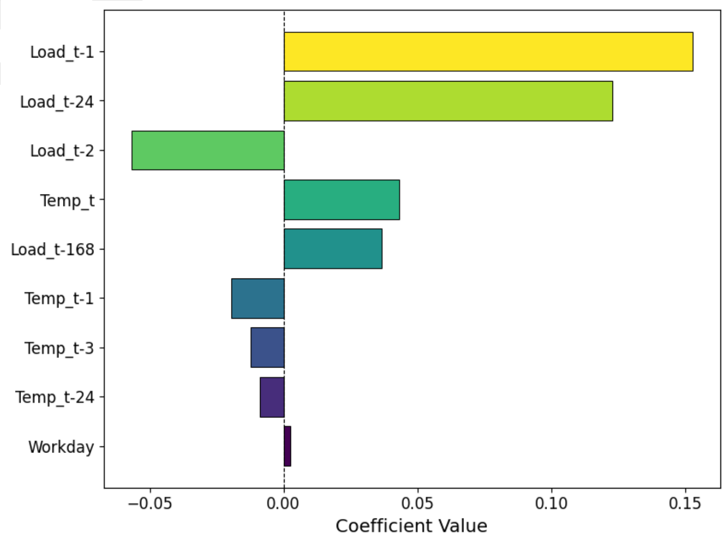
Analyzing the coefficients of the Ridge model shows that the most important predictor of electricity consumption is the consumption from one hour ago, $Load_{t-1}$. This result is fully consistent with the autocorrelation function analysis, which showed a strong short-term dependency. In addition, the significant coefficient of the 24-hour lag, $Load_{t-24}$ indicates the prominent role of daily patterns in determining electricity consumption. The presence of a positive coefficient for the 168-hour lag of electricity consumption $Load_{t-168}$ also indicates the influence of weekly cycles and confirms the findings from initial exploratory analyses. The negative coefficient of the $Load_{t-2}$ can be interpreted

10.22060/EEJ.2026.25648.5974

as a short-term corrective effect in response to sudden fluctuations in consumption. Regarding the temperature variable, it is observed that the effect of instantaneous temperature and temperature lags, although smaller than the consumption lags, is non-zero and stable. This indicates that temperature plays a secondary but meaningful role in determining electricity consumption. The negative coefficients of some temperature lags indicate the complex and non-linear nature of the effect of temperature on electricity consumption, which can be modeled by non-linear models.

Fig. 6 shows the coefficients of the Ridge Regression model. It is observed that the short-term and daily lags of electricity consumption contribute the most to predicting consumption, while the temperature lag variables have smaller but non-zero coefficients, indicating the secondary and delayed role of temperature in the linear model. Since Ridge Regression is only capable of modeling linear relationships, the small coefficients of temperature in this model do not necessarily mean that temperature is unimportant, but rather are due to the non-linear nature of the effect of temperature on electricity consumption. This observation motivates using non-linear models and advanced interpretability tools such as SHAP to examine the role of temperature in more detail.

Finally, the feature related to workday had a very small coefficient, indicating that the major pattern of differences between workdays and non-workdays is implicitly reflected in the temporal lags of electricity consumption.



10.22060/EEJ.2026.25648.5974

Fig. 6. Coefficients of the Ridge Regression model.

4-2- Random Forest model

The Random Forest model was trained with 200 trees without depth limitation. This setting allows the model to learn non-linear relationships and interactions between features, while a separate test set controls the risk of overfitting. Evaluating the model's performance showed that the Random Forest Regression model was able to predict a significant portion of the variation in electricity consumption. In the training set, MAE was 0.0088, RMSE was 0.0141, and the R^2 was 0.9971. In the test data, MAE = 0.0250, RMSE = 0.0354, and R^2 = 0.9770. The proximity of the error values and R^2 between training and testing indicates good generalization of the model, and this value is higher than the R^2 value of the Ridge Regression model (0.9658) with selected lags, indicating the ability of the non-linear model to identify complex effects and interactions between features.

Feature importance analysis for the Random Forest model shows that, similar to the Ridge Regression findings, short-term and daily electricity consumption lags are the most important features for prediction. However, the Random Forest model was able to identify the non-linear effect of some temperature lags. These features, which had a small coefficient in the linear model, gained significant relative importance in the Random Forest, indicating the non-linear and partial but meaningful role of temperature in determining electricity consumption. Fig. 7 clearly shows that the largest contribution to prediction comes from the 24, 1, and 168-hour lags of electricity consumption, and the effect of temperature is considered non-linearly in the model.

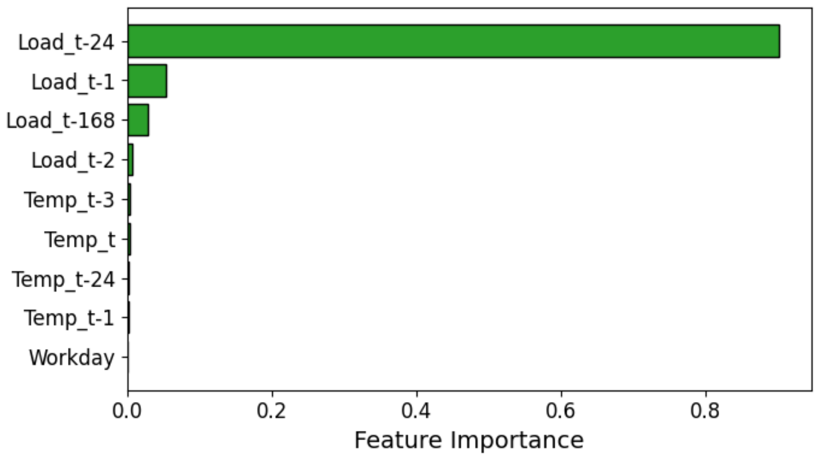


Fig. 7. Feature importance for the Random Forest model.

10.22060/EEJ.2026.25648.5974

Comparing with Ridge Regression shows that the linear model can identify the major part of the linear and short-term dependencies of consumption, but does not consider the non-linear effect of temperature and complex interactions between lags. In contrast, the Random Forest, by identifying the effect of temperature and complex interactions, was able to significantly increase the prediction accuracy. These findings show that in electricity consumption time series, where both short-term dependency and the effect of non-linear environmental variables are important, the use of non-linear models such as Random Forest is scientifically and practically very useful.

Furthermore, the feature importance analysis showed that some temperature lags, especially the temperature with a 3-hour delay, had higher importance than the temperature at the same hour. This finding clearly demonstrates the non-linear and delayed effect of temperature on electricity consumption. One scientific reason for this behavior is that electricity consumption is often dependent on the response of building heating and cooling systems, which operate with a time delay relative to temperature changes; therefore, past temperature is a better indicator of actual energy needs than instantaneous temperature. Also, people usually show their consumption behavior with a time delay, as ventilation or heating systems respond with a short delay. Consequently, the greater importance of the temperature with a 3-hour delay compared to the temperature at the same hour indicates the existence of delayed physical and behavioral processes that affect electricity consumption, and non-linear models can detect them. This analysis shows that in electricity consumption time series, using temperature features with different lags to identify real environmental effects is essential, and non-linear models are a suitable tool for this purpose.

4-3- XGBoost model

The performance of the XGBoost model in the training set was MAE = 0.0126, RMSE = 0.0167, and $R^2 = 0.9960$, and in the test set, it was MAE = 0.0244, RMSE = 0.0345, and $R^2 = 0.9781$. These results show that the model is both able to accurately learn the fluctuations in electricity consumption and maintain good generalization to new data, as the difference between the training and test sets is very small. Comparing this model with the previous models in Table 3 shows that:

- Ridge Regression model: (Test $R^2 = 0.9658$) As a baseline linear model, it identifies the major part of the linear and short-term dependencies but does not well represent the non-linear effect of temperature and interactions between features.
- Random Forest model: (Test $R^2 = 0.9770$) As a non-linear model with the ability to identify non-linear effects and interactions between lags, it was able to increase prediction accuracy and identify the effect of temperature and temperature lags partially but meaningfully.
- XGBoost model: (Test $R^2 = 0.9781$) This model was able to identify the non-linear effect and complex interactions between load and temperature even more accurately and had better generalization.

Table 3. Comparison of linear and non-linear models.

	Train			Test		
Method	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
XGBoost	0.0126	0.0167	0.9960	0.0244	0.0345	0.9781
Random Forest	0.0088	0.0141	0.9971	0.0250	0.0354	0.9770
Ridge Regression	0.0322	0.0472	0.968	0.0313	0.0431	0.9658

Generally, comparing the performance of the models shows that the simple linear model, which includes only the consumption of the previous hour and the temperature at the same hour, covers the short-term dependency and the effect of instantaneous temperature with $R^2 = 0.8735$, but information about periodic lags and non-linear effects is ignored in it. Adding selected load lags to the baseline Ridge Regression model increased the accuracy to $R^2 = 0.9658$, and the effect of short-term and daily lags was well identified, but the non-linear effect of temperature remained limited. The non-linear Random Forest model with $R^2 = 0.9770$ was able to identify non-linear effects and complex interactions between features in addition to load lags, creating a significant improvement over Ridge Regression.

Although the XGBoost and Random Forest models demonstrate very high predictive accuracy, a small gap between training and testing performance (approximately 1.8–2% in terms of R^2) can be observed. This behavior indicates a moderate level of overfitting, which is a known characteristic of high-capacity ensemble models when modeling complex electricity consumption patterns. While the models effectively capture the dominant structures of the dataset, this moderate overfitting suggests that their

10.22060/EEJ.2026.25648.5974

generalization capability may decrease when applied to buildings with substantially different consumption characteristics.

4-4- Residual Diagnostic Analysis

In addition to evaluating predictive accuracy, examining the statistical properties of model residuals provides further insight into the reliability and stability of forecasting models. Residual diagnostics help determine whether the model has adequately captured the underlying structure of the data and whether systematic patterns remain unexplained. In time-series forecasting, such analysis is particularly important because unmodeled temporal dependencies, heteroscedasticity, or distributional irregularities in the residuals may indicate limitations in the model structure.

To assess the adequacy of the best-performing model, a residual diagnostic analysis was conducted for the XGBoost model using several commonly applied statistical tests. These tests examine different aspects of residual behavior, including autocorrelation, distributional properties, and variance stability. The results of these diagnostic tests are summarized in Table 4.

The Durbin–Watson statistic was obtained as 1.6553, which lies within the acceptable range of 1.5 to 2.5 and indicates that there is no strong first-order autocorrelation in the residuals. However, the Ljung–Box test at lag 10 was statistically significant, suggesting that some higher-order dependencies remain in the residuals. This behavior is commonly observed in electricity load data due to the dynamic nature of consumption patterns, variations in user behavior, and the influence of external factors. The Jarque–Bera test also indicates that the residuals deviate from a normal distribution, which is likely associated with extreme values and sudden load fluctuations. In addition, the Breusch–Pagan test provides evidence of heteroscedasticity in the residuals.

Despite these characteristics, the mean of the residuals was very close to zero (0.0026), indicating that the model predictions do not exhibit significant systematic bias. Overall, although some classical statistical assumptions are not fully satisfied, the diagnostic results suggest that the XGBoost model successfully captures the major structure of the data and maintains stable and reliable predictive performance.

Table 4. Residual diagnostic statistics for the XGBoost model.

10.22060/EEJ.2026.25648.5974

Test	Statistic	P-Value	Conclusion
Durbin-Watson	1.6553	N/A	No strong first-order autocorrelation
Ljung-Box, lag 10	208.37	< 0.001	Evidence of higher-order residual dependence
Jarque-Bera	3374.64	< 0.001	Non-normal residual distribution
Breusch-Pagan	9.425	0.002	Evidence of heteroscedasticity

4-5- Comparative Performance Evaluation and Statistical Significance Analysis

To provide a comprehensive and reliable comparison of the developed forecasting models, their predictive performance was evaluated on the test dataset using standard error metrics, including MAE, RMSE, and the coefficient of determination (R^2). In addition to point estimates, 95% confidence intervals were calculated using the Circular Block Bootstrap method in order to assess the robustness and statistical stability of the model performances under temporal dependency. Furthermore, to examine whether the observed differences between the best-performing models were statistically significant, the Diebold–Mariano (DM) test was conducted on their forecasting errors. This comparative analysis makes it possible to evaluate not only the numerical accuracy of the models but also the statistical reliability of the observed performance differences.

Table 5 presents the evaluation results of the developed models on the test dataset. The confidence intervals were calculated using the Circular Block Bootstrap method with a block size of 24 hours. The results indicate that all three models demonstrate very strong performance in electricity load forecasting. Among them, the XGBoost model achieved the highest predictive accuracy with an R^2 value of 0.9781 and the lowest forecasting error (MAE = 0.0244), making it the most accurate model in this study.

An important aspect of the results is the reporting of 95% confidence intervals for all performance metrics. The relatively narrow width of these intervals indicates a high degree of robustness in the model predictions despite the presence of temporal fluctuations in the time-series data. For example, the confidence interval width for the R^2 metric in the XGBoost model is approximately 0.0085, reflecting stable model performance under resampling conditions.

A comparison of the confidence intervals further reveals that the lower bound of the confidence intervals for the non-linear models (Random Forest and XGBoost) lies entirely above the upper bound of the confidence interval for the linear Ridge Regression model. This observation provides strong statistical evidence that the performance improvement achieved by the more complex models is not the result of random variation or sampling effects in the test dataset. In addition, the proximity of the point estimates to the center of their respective confidence intervals suggests that the training procedure is well balanced and that no abnormal overfitting behavior is present.

Table 5. Comparison of the performance of the proposed models along with the 95% confidence interval.

Model	MAE [95% CI]	RMSE [95% CI]	R^2 [95% CI]
Ridge Regression	0.0313 [0.0286, 0.0344]	0.0431 [0.0385, 0.0482]	0.9658 [0.9592, 0.9715]
Random Forest	0.0250 [0.0225, 0.0276]	0.0354 [0.0314, 0.0399]	0.9770 [0.9720, 0.9813]
XGBoost	0.0244 [0.0222, 0.0268]	0.0345 [0.0307, 0.0389]	0.9781 [0.9735, 0.9820]

The relatively small width of the confidence intervals confirms the stability of model performance against data variability and indicates that the predictive accuracy remains consistent even in the presence of temporal dependency within the dataset.

To further examine the superiority of the XGBoost model over the Random Forest model, the Diebold–Mariano (DM) test was applied to the prediction residuals of the two models. Although the numerical difference in the R^2 metric between the models (0.0011) appears small, the DM test provides a more rigorous statistical comparison of forecasting accuracy. The results of the DM test produced a statistic value of 2.1487 and a p-value of 0.0317. Since the p-value is below the commonly used significance threshold of 0.05, the null hypothesis of equal predictive accuracy between the two models is rejected. This finding indicates that the observed improvement in forecasting performance achieved by the XGBoost model is statistically significant rather than the result of random fluctuations in the evaluation dataset.

Therefore, based on both predictive accuracy metrics and statistical significance analysis, the XGBoost model can be considered the most reliable and stable forecasting model among the evaluated approaches for the electricity load prediction task in this study. This result highlights the effectiveness of gradient boosting methods in capturing complex nonlinear relationships and temporal patterns in electricity consumption data.

4-6- Comparison with Classical Time-Series Baseline Models

10.22060/EEJ.2026.25648.5974

To further evaluate the effectiveness of the proposed machine learning models, their performance was compared with classical time-series forecasting approaches commonly used in energy demand prediction. Specifically, the ARIMA and SARIMAX models were implemented as baseline methods. These models represent traditional linear autoregressive frameworks that have been widely applied in electricity load forecasting. By comparing their performance with the proposed machine learning models, it becomes possible to assess whether the additional complexity of non-linear algorithms is justified for the building energy consumption forecasting problem considered in this study.

To assess the necessity of using non-linear machine learning models, two classical time-series models, ARIMA and SARIMAX, were implemented as baselines. The univariate ARIMA model with the best structure, ARIMA(2,1,2) was selected based on the lowest AIC value. This model achieved an R^2 value of 0.058 on the test data, indicating the inability of linear autoregressive structures to model the complex energy consumption patterns of the studied building.

In the next step, the SARIMAX model was executed by incorporating a daily seasonal structure ($m = 24$) and the exogenous temperature variable. This model was able to significantly improve performance and achieved an R^2 value of 0.8859 on the test data. This performance improvement indicates that part of the energy consumption dynamics is driven by seasonality and external factors. Nevertheless, the proposed non-linear machine learning models still delivered substantially better performance. The XGBoost model recorded an R^2 value of 0.9781, and the Random Forest achieved an R^2 value of 0.9770, representing the highest prediction accuracy. The approximately 9% difference between SARIMAX and the non-linear models shows that linear structures—even when enhanced with seasonality and exogenous inputs—are unable to fully capture the complex and non-linear relationships present in building energy consumption data.

The results presented in Table 6 show that transitioning from simple linear models (ARIMA) to more advanced linear models incorporating seasonality and exogenous variables (SARIMAX) leads to a significant increase in prediction accuracy. However, even the SARIMAX model could not approach the performance of the non-linear tree-based models. This performance gap indicates that building energy consumption is influenced by complex and non-linear interactions among variables such as

10.22060/EEJ.2026.25648.5974

temperature, temporal patterns, user behavior, and environmental conditions. Tree-based models like XGBoost, due to their ability to model high-order interactions and non-linear structures, are capable of extracting these hidden patterns, whereas linear autoregressive models—even with seasonality—suffer from structural limitations. Therefore, the use of non-linear machine learning models in this research has been not only a methodological choice but also an empirical necessity for achieving high accuracy in building energy consumption forecasting.

Table 6. Comparison of model performance on the test data

Model	Model Type	Test R^2	RMSE
ARIMA (2,1,2)	Univariate linear	0.058	0.2266
SARIMAX (1,1,1)(1,0,1,24)	Seasonal linear + exogenous variable	0.8859	0.0789
Ridge	Multivariate linear	0.9658	0.0431
Random Forest	Nonlinear (Ensemble)	0.9770	0.0354
XGBoost	Nonlinear (Ensemble)	0.9781	0.0345

4-7- SHAP and LIME: Global and Local Interpretability Analysis

To analyze the behavior of the XGBoost model and examine the impact of input features on consumption prediction, SHAP and LIME methods were used. Both methods enable the analysis of feature importance for complex models at global and local levels. The SHAP analysis of the XGBoost model shows that, similar to the Ridge Regression and Random Forest models, the greatest impact on consumption prediction belongs to the short-term and daily lags of electricity consumption, but temperature lags with a delay of 1–3 hours also have a significant role, indicating the delayed effect of heating and cooling systems and consumer behavior. SHAP plots in Fig. 8 show the direction and magnitude of the impact of each feature, and enable a detailed analysis of the non-linear effect of temperature and interactions between lags. The XGBoost model with $R^2 = 0.9781$ provided better performance than the Random Forest, and its SHAP analysis enabled a more accurate interpretation of the non-linear effect of temperature and lags, so that the direction and true importance of each feature were clearly identified. These results show that for predicting electricity consumption in time series with short-term, daily, and environmental effects, using non-linear models along with SHAP analysis can provide a suitable scientific explanation regarding the importance and direction of the effect of features.

10.22060/EEJ.2026.25648.5974

To provide a meaningful and non-random interpretation of the SHAP waterfall plots, three data samples were selected based on the level of electricity consumption. Specifically, one sample corresponding to peak load, one from typical load, and one from low load were analyzed. This approach allows for examining the model's behavior in different consumption regimes and thus prevents bias resulting from arbitrary sample selection.

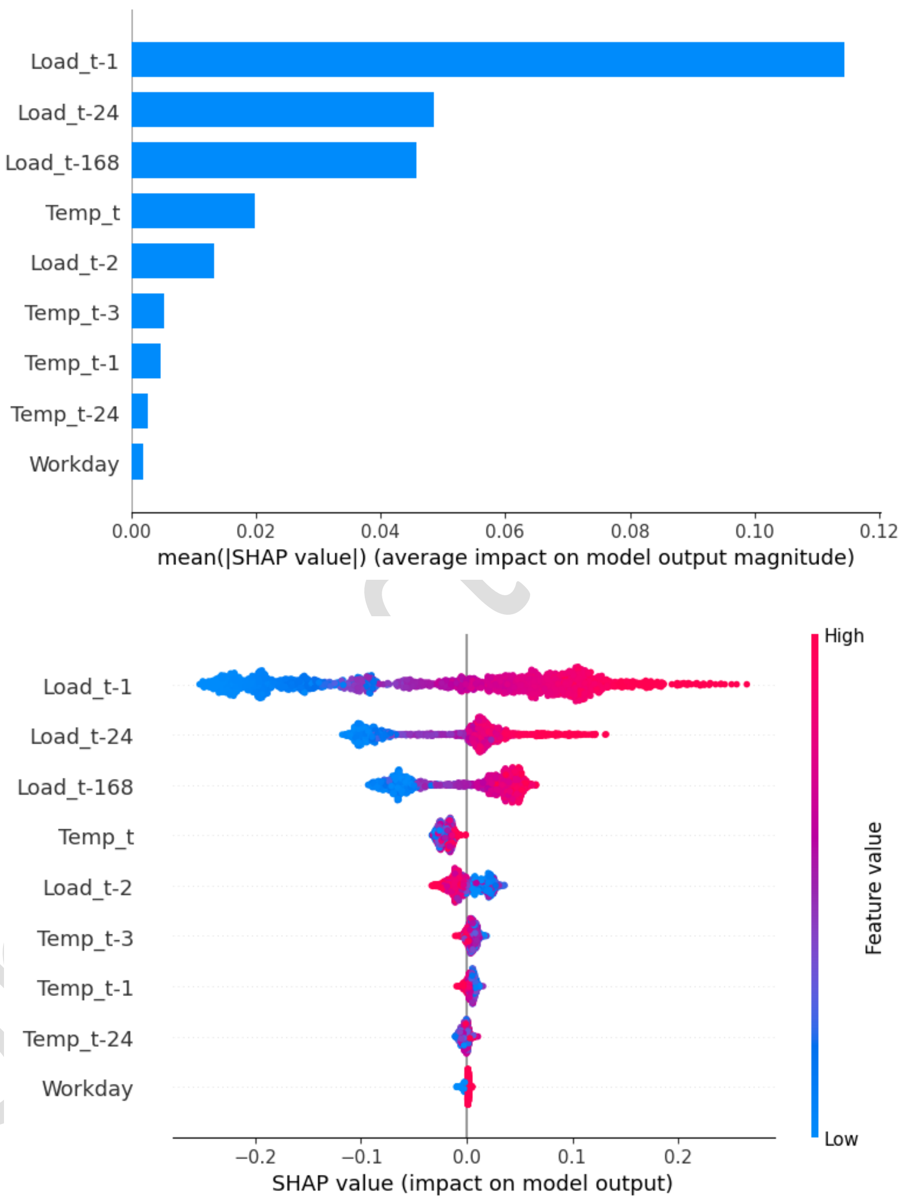
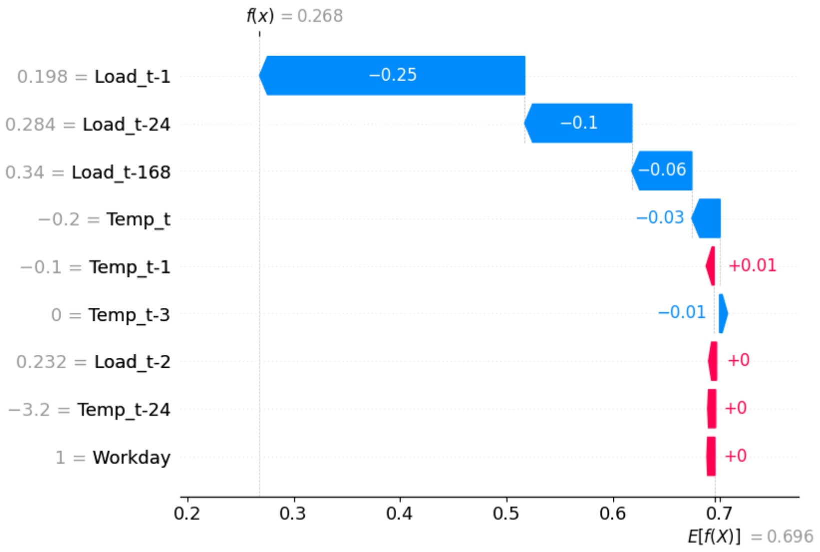


Fig. 8. SHAP plot and feature importance for the XGBoost model.

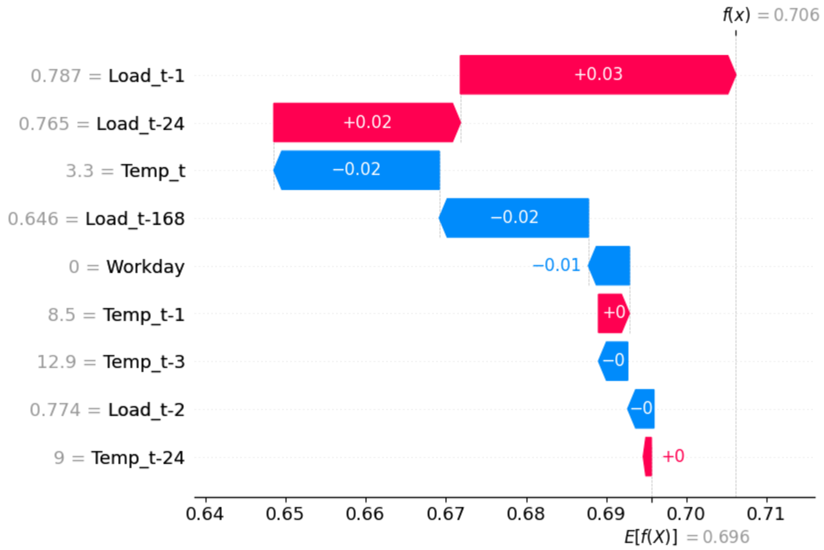
Fig. 9, using explainability waterfall plots, compares three distinct situations of low load, typical load, and peak load for the XGBoost model, and clearly shows how changes in past consumption patterns shift the model's output around the base value E[f(X)] = 0.696. In the low load (Fig. 9-a), the model's

10.22060/EEJ.2026.25648.5974

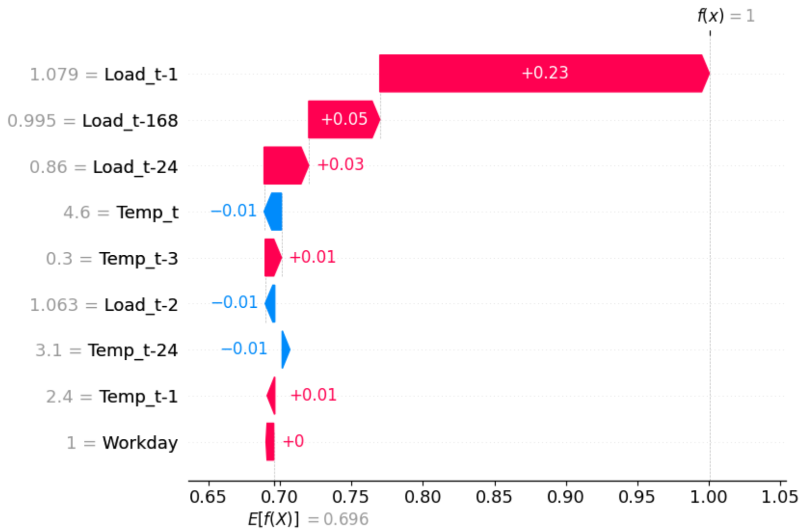
predicted value f(X) = 0.268, which is significantly lower than the base value; in this case, electricity consumption in the past hours are all at low levels (e.g., $Load_{t-1} \approx 0.20$, $Load_{t-24} \approx 0.28$ and $Load_{t-168} \approx 0.34$) and each has a significant negative effect (approximately -0.25, -0.10 and -0.06) on reducing the model's output, so that the sum of these negative effects is the main factor in forming the low-load forecast, while the effect of temperature variables and day type (workday or non-workday) remains very small and almost neutral. This situation shows that the model interprets the simultaneous decrease in short-term, daily and weekly consumption as a strong indicator of a low-load system. In the typical load (Fig. 9-b), the model's predicted value f(X) = 0.706, which is very close to the base value; in this sample, the consumption one hour ago at a moderate level ($Load_{t-1} \approx 0.79$) creates a positive effect, although limited (about +0.03), but this effect is largely neutralized by the negative effects of consumption with longer delays and lower temperatures (such as $Temp_t \approx 3.3$ with an effect of -0.02). Analyzing this plot indicates a kind of balance between increasing and decreasing forces that leads to a prediction of consumption at a moderate level and shows that the model is not facing strong indicators of peak consumption or low-load in this case. Finally, in the peak load (Fig. 9-c), the model's predicted value reaches f(X) = 1, which is the highest output level; in this case, electricity consumption in the past hours is very high, especially $Load_{t-1} \approx 1.08$, which alone contributes a large positive share of about +0.23 to increasing the output, and daily and weekly consumption also reinforce this trend with smaller positive effects (about +0.03 and +0.05), while the effect of temperature variables remains very limited and corrective.


f(x) = 0.268
0.198 = Load_t-1
0.284 = Load_t-24
0.34 = Load_t-168
−0.2 = Temp_t
−0.1 = Temp_t-1
0 = Temp_t-3
0.232 = Load_t-2
−3.2 = Temp_t-24
1 = Workday
−0.25
−0.1
−0.06
−0.03
+0.01
−0.01
+0
+0
+0
0.2
0.3
0.4
0.5
0.6
0.7
E[f(X)] = 0.696

(a)


f(x) = 0.706
0.787 = Load_t-1
0.765 = Load_t-24
3.3 = Temp_t
0.646 = Load_t-168
0 = Workday
8.5 = Temp_t-1
12.9 = Temp_t-3
0.774 = Load_t-2
9 = Temp_t-24
+0.03
+0.02
−0.02
−0.02
−0.01
+0
−0
−0
+0
0.64
0.65
0.66
0.67
0.68
0.69
0.70
0.71
E[f(X)] = 0.696

(b)


f(x) = 1
1.079 = Load_t-1
0.995 = Load_t-168
0.86 = Load_t-24
4.6 = Temp_t
0.3 = Temp_t-3
1.063 = Load_t-2
3.1 = Temp_t-24
2.4 = Temp_t-1
1 = Workday
+0.23
+0.05
+0.03
−0.01
+0.01
−0.01
−0.01
+0.01
+0
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05
E[f(X)] = 0.696

(c)

10.22060/EEJ.2026.25648.5974

Fig. 9. SHAP waterfall plots. a) Low load, b) Typical load, c) Peak load.

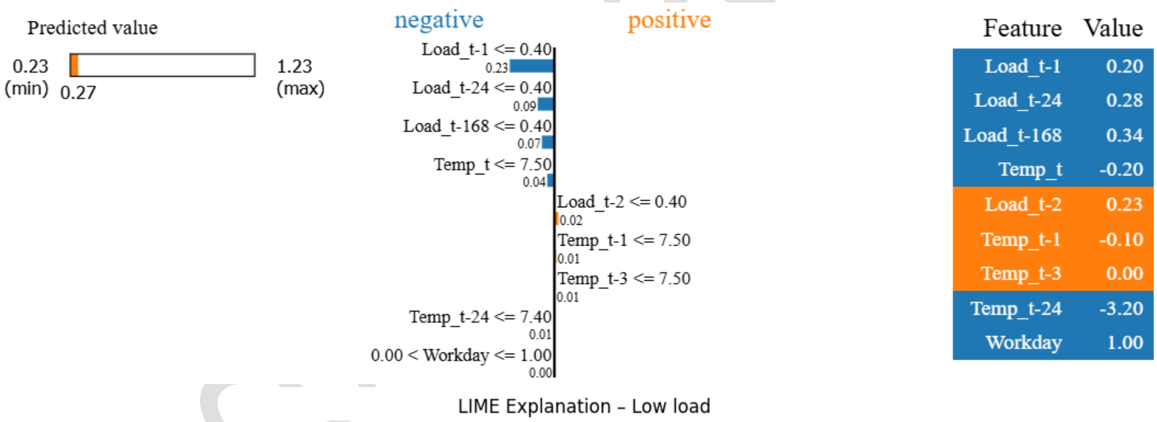
Overall, the waterfall analysis of these three data points shows that the model's transition from low load to typical load and then to peak load is mainly due to the change in sign and intensity of the effect of past consumption, while temperature variables and day type have played a secondary role in all three situations and have acted more as minor modifying factors around the dominant consumption pattern.

Fig. 10 shows the results of data analysis using the LIME method to explain the XGBoost model's predictability in forecasting peak load conditions. In this diagram, the goal is to show how the model relies on past electricity consumption values (such as consumption one hour ago, 24 hours ago, and 168 hours ago) as well as temperature variables and day type to predict the peak load amount for a specific hour. The left side of the figure shows the contribution of each feature to predictions of low load, typical load, and peak load; green bars indicate the positive effect and red bars indicate the negative effect of that feature on increasing or decreasing the predicted load amount. This figure shows how hourly consumption patterns throughout the year and environmental conditions play a role in predicting peak electricity load and help to better understand the model's behavior on real data.

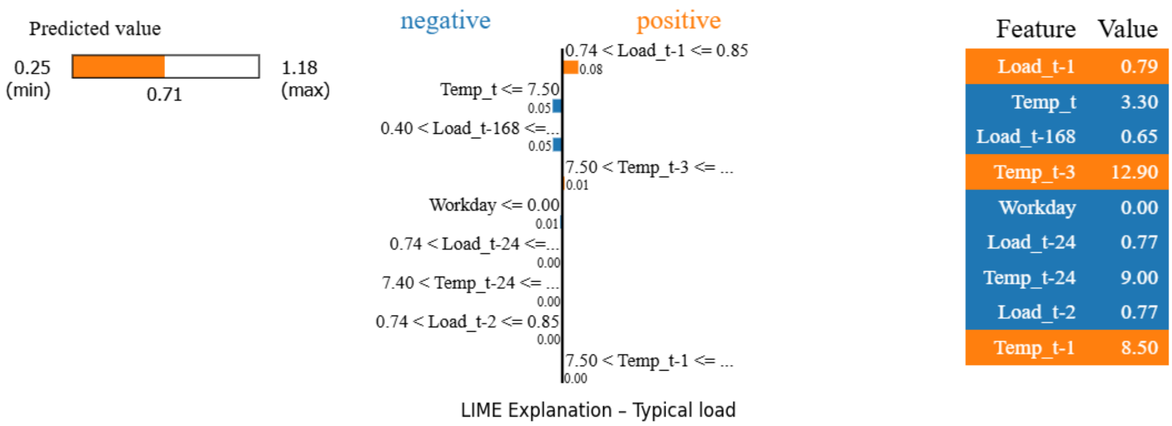
Fig. 10-a shows the output of the LIME method for explaining the XGBoost model's prediction in the low load, where the model's predicted value is around 0.27 and is close to the minimum model output and much lower than the typical load. In this sample, electricity consumption in the past hours is all low; specifically, $Load_{t-1} = 0.20$, $Load_{t-24} = 0.28$, and $Load_{t-168} = 0.34$, all of which are in the range ≤ 0.40, and all three have had a significant negative contribution to the prediction. This shows that the model consistently interprets the decrease in short-term, daily, and weekly consumption as a strong indicator of a low-load. In contrast, some temperature variables such as $Temp_t = -0.20$ and $Temp_{t-1} = -0.10$, have had very small positive effects, but these effects are negligible compared to the negative impact of past consumption. The day type feature also remained almost without effect. Fig. 10-b shows the output of the LIME method for explaining the model's prediction in the typical load, where the model's predicted value is around 0.71, which is between the minimum and maximum model outputs (approximately 0.25 to 1.18). The most important increasing factor in this sample is the variable $Load_{t-1}$ in the range of 0.74 to 0.85 with an actual value of 0.79, which has the largest positive

10.22060/EEJ.2026.25648.5974

contribution to increasing the prediction and shows that consumption one hour ago, even at an average level, plays a major role in determining current consumption. In contrast, the current temperature of 3.30 degrees Celsius, which is in the range of ≤ 7.5, has had a significant negative effect on the prediction, causing a decrease in the output amount. Also, consumption 168 hours ago with a value of 0.65 and some temperature variables with different time delays, such as $Temp_{t-3} = 12.9$ and $Temp_{t-1} = 8.5$, have had smaller effects. The effect of the day type feature was also negligible, indicating that day type did not play a decisive role in this data point. Overall, in the typical load, the balance between the positive effect of short-term consumption and the negative effect of temperature leads to a prediction of an intermediate value. Fig. 10-c shows the output of the LIME method for explaining the model's prediction in the peak load, and as can be seen, electricity consumption in the past hours, especially consumption one hour ago and one day ago, has had the greatest positive impact on predicting the peak load, while some temperature conditions have a reducing effect.

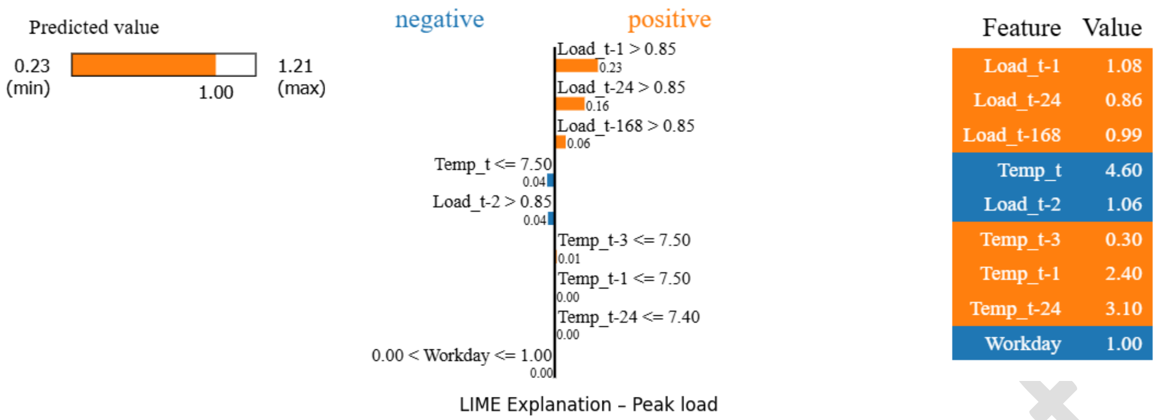


(a)



(b)

10.22060/EEJ.2026.25648.5974



(c)

Fig. 10. Local explainability with LIME. a) Low load, b) Typical load, c) Peak load.

Comparing the LIME outputs for the peak load, typical load, and low load shows that the model's behavior in all three states follows a consistent decision-making logic, but the intensity and direction of feature effects change with the consumption level. In the peak load, the model's predicted value is equal to 1 and close to the maximum output; in this situation, electricity consumption in the past hours, especially $Load_{t-1}$ with a very high value (more than 1 on the model scale), has a dominant and determining share in increasing the output, and other delayed daily and weekly consumptions also create reinforcing positive effects, so that the effect of temperature variables and day type is very insignificant compared to past consumption. In contrast, in the typical load, although consumption one hour ago remains the most important increasing factor, its value is in an intermediate range, and at the same time, the negative effect of low temperature causes the model output to reach a moderate level (around 0.71); in this case, the model's decision is the result of a balance between the positive effects of short-term consumption and the negative effects of temperature conditions. Finally, in the low load, the direction of the dominant effect is completely reversed; past consumptions are all at low levels, and all three short-term, daily, and weekly components have significant negative effects on the prediction simultaneously, driving the model output towards the minimum (around 0.27), while the effects of temperature and day type remain small and corrective. This comparison clearly shows that the consumption prediction model, regardless of the consumption level, relies most on the temporal patterns of electricity consumption, and the change between peak load, typical load, and low load is mainly due to changes in the values of consumption in past hours, while environmental variables only play a modulating role around this dominant pattern.

The results of this study show that combining non-linear machine learning models with advanced interpretability methods provides a powerful framework for predicting and analyzing hourly electricity consumption. Comparing the performance of the models showed that while Ridge Regression is able to well identify short-term linear dependencies and main periodic patterns, it is limited in modeling non-linear effects and complex interactions between consumption and temperature. In contrast, Random Forest and XGBoost models, by leveraging non-linear structures and learning feature interactions, provided a significant improvement in prediction accuracy, which was particularly evident in the increase in the R^2 in the test dataset.

5. Limitations and Future Research Directions

Although the proposed framework in this research has demonstrated desirable forecasting performance and appropriate interpretability, several important limitations should be considered. First, the analyses performed are based on data from a single commercial building; therefore, the obtained quantitative results may be influenced by the operational characteristics and usage patterns specific to that building. Second, the dataset used covers only one year of data from a specific climatic region and, thus, may not reflect all the variations present in weather conditions and consumption patterns observed in different geographic regions. Third, this study examined only one type of building usage, whereas residential, industrial, educational, and office buildings can exhibit different consumption behaviors, temporal dependencies, and sensitivities to climatic variables.

Accordingly, the results of this research should primarily be regarded as a validation of the proposed interpretable framework for electricity load forecasting, rather than as a fully generalizable prediction model for all buildings and climatic conditions. In future research, evaluating this framework using different types of buildings, diverse climatic regions, longer time periods, and data related to different geographic areas can provide a more comprehensive picture of the stability, transferability, and scalability of the proposed approach.

10.22060/EEJ.2026.25648.5974

Although the present study has focused on a commercial building located in a specific climatic region, the structure of the proposed framework possesses sufficient flexibility for application in other types of buildings and geographic regions. Therefore, it is suggested that future research investigate the transferability of this framework across different climates and for buildings with various usages, using larger and more diverse datasets.

6. Conclusion

In this study, a comprehensive framework for predicting hourly electricity consumption was presented using linear and nonlinear models and advanced interpretability methods. The results showed that while linear models such as Ridge Regression can identify short-term dependencies and main periodic patterns, they are limited in modeling the nonlinear effects of temperature and complex interactions between variables. In contrast, Random Forest and XGBoost models, by achieving higher prediction accuracy, were able to extract more realistic behavioral and physical patterns of electricity consumption. Feature importance analysis for the Random Forest model showed that short-term and daily lags of electricity consumption are still the most important predictive factors, but unlike linear models, temperature lags with a time delay play a more prominent role. Specifically, it was observed that temperature with a delay of 1 to 3 hours is often more important than the temperature at the same hour. This behavior can be attributed to the thermal inertia of buildings and the gradual reaction of heating and cooling systems, which causes the effect of temperature on electricity consumption to appear with a time delay. Such patterns are difficult to identify by linear models and highlight the advantage of tree-based and boosted models in modeling real physical-behavioral phenomena. The XGBoost model, in addition to providing a slight improvement in prediction accuracy compared to the Random Forest, enabled more accurate interpretable analysis through SHAP. SHAP analysis showed that the effect of temperature on electricity consumption is highly non-linear and changes depending on the consumption level and the value of load lags. These results show that the XGBoost model is able to implicitly identify different consumption regimes and adapt the model's behavior to real conditions.

SHAP- and LIME-based analyses showed that short-term and daily lags of electricity consumption are the most important predictive factors, while the effect of temperature appears mainly non-linearly and

10.22060/EEJ.2026.25648.5974

with a time delay, reflecting thermal inertia and adaptive consumer behavior. To complete the model interpretability analysis, SHAP waterfall plots were used for three data samples, including peak load, typical load, and low load conditions. This local analysis showed that the contribution of each feature to the final prediction is highly dependent on the instantaneous conditions of the system, and the direction of the effect of some variables (such as temperature) can change between samples. In addition, the use of LIME as a complementary and model-agnostic method confirmed the stability of the local interpretations obtained from SHAP. The LIME results for the selected samples were consistent with the SHAP findings and showed that the model's decision-making at the level of individual samples is also based on logical and physical justifications.

The main innovation of this study is the presentation of an integrated framework in which advanced non-linear modeling is combined with multi-layered interpretability analysis. Unlike many previous studies that focus solely on improving accuracy, this study shows how to achieve both high accuracy and a deep understanding of the role of load lags and temperature. The combination of global (Feature Importance and SHAP summary) and local (SHAP waterfall and LIME) analyses enables the extraction of reliable and interpretable knowledge from the model, which can be directly used in load management, energy planning, and data-driven policymaking. The combination of global and local interpretability analyses not only increased the transparency of model decision-making but also strengthened its reliability for practical applications. These results show that the simultaneous use of non-linear models and interpretable tools can provide an effective and reliable approach to predicting electricity consumption and supporting intelligent decision-making in future energy systems.

Although the present study has focused on a commercial building located in a specific climatic region, the proposed framework possesses sufficient flexibility and adaptability for application in other types of buildings and geographic regions. However, to more accurately evaluate the generalizability and transferability of this framework, it is necessary for future research to examine its performance across different climates, for buildings with various usages, and using larger and more diverse datasets. Such studies can provide a more comprehensive understanding of the stability, transferability, and efficiency of the proposed framework.

References

[1] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, International journal of forecasting, 30(4) (2014) 1030–1081.
[2] J.W. Taylor, Short-term load forecasting with exponentially weighted methods, IEEE transactions on Power Systems, 27(1) (2011) 458–464.
[3] T. Hong, S. Fan, Probabilistic electric load forecasting: A tutorial review, International Journal of Forecasting, 32(3) (2016) 914–938.
[4] J.W. Taylor, Short-term electricity demand forecasting using double seasonal exponential smoothing, Journal of the Operational Research Society, 54(8) (2003) 799–805.
[5] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.
[6] R.H. Shumway, D.S. Stoffer, Time series analysis and applications, 2016.
[7] J. Wang, H. Wang, Interpretable models for energy load forecasting: Challenges and opportunities, Energy Reports, 7 (2021) 214–230.
[8] H.S. Hippert, C.E. Pedreira, R.C. Souza, Neural networks for short-term load forecasting: A review and evaluation, IEEE Transactions on power systems, 16(1) (2001) 44–55.
[9] R. Mathumitha, P. Rathika, K. Manimala, Intelligent deep learning techniques for energy consumption forecasting in smart buildings: a review, Artificial Intelligence Review, 57(2) (2024) 35.
[10] L. Breiman, Random forests. Machine learning, 45(1) (2001) 5–32.
[11] T. Chen, XGBoost: A Scalable Tree Boosting System, Cornell University, 2016.
[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR), 51(5) (2018) 1–42.
[13] E.L. Alba, G.A. Oliveira, M.H.D.M. Ribeiro, É.O. Rodrigues, Electricity consumption forecasting: An approach using cooperative ensemble learning with SHapley additive explanations, Forecasting, 6(3) (2024) 839–863.
[14] A.M. Salih, Z. Raisi-Estabragh, I.B. Galazzo, P. Radeva, S.E. Petersen, K. Lekadir, G. Menegaz, A perspective on explainable artificial intelligence methods: SHAP and LIME, Advanced Intelligent Systems, 7(1) (2025) 2400304.
[15] B. Teixeira, T. Pinto, Z. Vale, Leveraging XAI Techniques for Context-Aware Energy Consumption Forecasting, in: World Conference on Explainable Artificial Intelligence 2025, pp. 205–217.
[16] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access, 6 (2018) 52138–52160.
[17] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems, 30 (2017).
[18] R.M. Rizk-Allah, L.M. Abouelmagd, A. Darwish, V. Snasel, A.E. Hassanien, Explainable AI and optimized solar power generation forecasting model based on environmental conditions, PloS one, 19(10) (2024) e0308002.
[19] D. Garreau, U. Luxburg, Explaining the explainer: A first theoretical analysis of LIME, in: International conference on artificial intelligence and statistics 2020, pp. 1287–1296.
[20] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European conference on machine learning and knowledge discovery in databases 2020, pp. 417–431.
[21] H. Qu, Z. Ye, Multi-objective optimization operation of micro energy network with energy storage system based on improved weighted fuzzy method, Energy reports, 9 (2023) 1995–2003.
[22] Y. Itani, M.R. Soliman, M. Kahil, Recovering energy by hydro-turbines application in water transmission pipelines: A case study west of Saudi Arabia, Energy, 211 (2020) 118613.
[23] K. Hagihara, C. Iba, S. Hokoi, Effective use of a ground-source heat-pump system in traditional Japanese "Kyo-machiya" residences during winter, Energy and Buildings, 128 (2016) 262–269.
[24] M. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, Proceedings of the ACM SIGKDD, (2016).
[25] J.N. Gopal, B. Madhu, S. Somu, A.J. Anand, Explainable AI (XAI) for energy demand forecasting, in: Neural Networks and Graph Models for Traffic and Energy Systems, IGI Global Scientific Publishing, 2025, pp. 115–138.
[26] J.N. Mupenzi, D. Witarsyah, A. Kusnadi, A.S. Sunge, An explainable data-driven framework for short-term residential energy forecasting, Energy and Buildings, (2026) 117713.
[27] Q. Wen, Y. Liu, Feature engineering and selection for prosumer electricity consumption and production forecasting: A comprehensive framework, Applied Energy, 381 (2025) 125176.
[28] A.M.E. Saleh, M. Arashi, B.G. Kibria, Theory of ridge regression estimation with applications, John Wiley & Sons, 2019.

[29] J.O.Q. Quispe, A.C.F. Quispe, N.C.L. Calvo, O.C. Toledo, Analysis and Selection of Multiple Machine Learning Methodologies in PyCaret for Monthly Electricity Consumption Demand Forecasting, Mater. Proc. , 18(5) (2024).