Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079734

ARTICLE

Month-Conditioned Boosting Framework with SHAP-in-the-Loop for Short-Term Electricity Load Forecasting

Jinsung Park[1,#], **Jaehyuk Lee**[1,2,#] **and Eunchan Kim**[1,3,*]

[1]Department of Information Systems, Hanyang University, Seoul, Republic of Korea
[2]Institute of IT Convergence Technology, Seoul National University of Science and Technology, Seoul, Republic of Korea
[3]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea

*Corresponding Author: Eunchan Kim. Email: eckim@hanyang.ac.kr

[#]These authors contributed equally to this work

Received: 27 January 2026; Accepted: 19 March 2026; Published: 08 May 2026

ABSTRACT: Accurate short-term load forecasting is essential for reliable power system operation, particularly under the increasing uncertainty caused by abnormal weather and socio-economic fluctuations. This study presents a month-conditioned boosting framework that integrates SHapley Additive Explanations (SHAPs) into model refinement. A baseline XGBoost model was first compared with linear and tree-based regressors, followed by enhancements through lagged and rolling-window features as well as loss weighting for vulnerable months. To further improve the performance, SHAP analysis was employed to identify the dominant error-contributing features, which guided the construction of targeted month-specific interaction terms for retraining. Experimental results based on rolling-origin cross-validation showed that this approach significantly reduced the RMSE and MAPE, particularly during high-variance summer months. Moreover, the SHAP interpretation revealed the varying roles of seasonal demand structures and socio-economic mobility, thereby enhancing transparency and operational insight. The proposed framework demonstrated that embedding explainability into the learning loop improved predictive accuracy and ensured interpretability, offering a data-driven solution for electricity demand forecasting in practical settings.

KEYWORDS: Electricity demand forecasting; explainable machine learning; feature engineering; SHAP analysis

1 Introduction

Monthly electricity load forecasting plays a critical role in energy system planning and operation, supporting decisions related to generation scheduling, fuel procurement, demand management, and infrastructure maintenance [1,2]. In recent years, the increasing frequency of abnormal weather events such as record-breaking heatwaves and unusually warm winters in Korea and other regions has increased the uncertainty of temperature-based forecasts [3,4]. These climatic anomalies, combined with meteorological, economic, and demographic dynamics, exacerbate the non-stationarity of monthly demand patterns. For instance, even under the same average temperature, electricity demand responses in August may differ substantially from those in January, and the relative importance of explanatory variables can shift seasonally [5,6]. This underscores the need for forecasting approaches that achieve high accuracy while providing transparent explanations of changes in predictive performance across different months.

Tree-based boosting models such as XGBoost remain highly competitive with deep learning for tabular datasets because of their robustness in capturing nonlinear relationships among heterogeneous covariates,

including meteorological, economic, and demographic factors [7–9]. When augmented with lag features and smoothing techniques, these models can deliver high predictive accuracy [10,11]. Moreover, tree-based models enable decision paths to be traced through split rules and feature contributions, and when integrated with SHapley Additive exPlanations (SHAP), they can quantitatively reveal the reasoning behind predictions, which enhances accountability in operational contexts. However, most existing literature has focused on large-scale deep learning models, applying SHAP mainly for post-hoc interpretation, which limits the use of explainability as an active driver of model improvement [12].

In contrast, herein, we propose a framework employing a closed-loop framework that uses model explanations directly to guide targeted retraining. First, a leakage-free baseline model is used to identify prediction-vulnerable months, and additional weights are applied to the training loss for these months to evaluate the performance improvements. For any month that still exhibits the highest error after this step, a SHAP importance analysis is performed to identify the most influential feature. Based on this feature, an interaction feature is generated and incorporated into the model, which is then retrained.

Using data from 2022 to 2023 for training and 2024 for evaluation, we optimized the model to capture contemporary demand patterns shaped by abnormal weather and seasonal variation. This strategy improved accuracy when most needed while controlling feature proliferation and maintaining interpretability. Detailed descriptions of the feature engineering process and the hyperparameter settings are provided in the Methods section.

The main contributions of this study are as follows:

First, unlike most previous studies where SHAP is used only for post-hoc interpretation, we propose a SHAP-in-the-loop framework in which model explanations are actively used to guide feature construction and retraining.

Second, the proposed method introduces a targeted month-conditioned interaction feature generation strategy, where the feature with the largest SHAP contribution in the highest-error month is used to construct an interpretable interaction variable.

Third, we design an error-driven retraining pipeline that combines error diagnosis, explainability analysis, and targeted feature engineering within a unified forecasting workflow.

2 Related Works

2.1 Statistical and Econometric Approaches

Electricity load forecasting has been extensively studied across multiple temporal horizons, ranging from minute-ahead very short-term forecasting to monthly and multi-year long-term planning. Early studies primarily relied on statistical and econometric models such as autoregressive integrated moving average (ARIMA) [13], exponential smoothing, and least absolute shrinkage and selection operator (LASSO) regression [14]. These methods have been widely adopted because of their transparency, well-established theoretical foundations, and ease of implementation.

However, traditional statistical approaches often struggle to capture nonlinear relationships among load drivers and the evolving interactions between meteorological, economic, and demographic variables [15]. To address such nonlinearities, machine learning-based methods such as feature selection with least squares support vector machines have also been explored for short-tern load forecasting [16]. These limitations become particularly pronounced in modern power systems, where demand patterns are increasingly affected by abnormal weather conditions, economic fluctuations, and structural changes in energy consumption behavior.

2.2 Deep Learning Approaches for Load Forecasting

To address the limitations of classical statistical models, a wide range of machine learning and deep learning approaches have been proposed. Neural network architectures such as deep neural networks (DNNs) [17], convolutional neural networks (CNNs), and recurrent neural networks (RNNs) [18] have demonstrated strong performance in short-term load forecasting tasks.

In particular, RNN-based architectures, including long short-term memory (LSTM) networks [19], gated recurrent unit (GRU) models [20], and temporal convolutional networks (TCNs) [21], are widely used to capture temporal dependencies in sequential load data. Hybrid architectures have also been proposed, such as CNN–GRU models with attention mechanisms [22] or decomposition-based frameworks combined with RNN or TCN modules [23], which further improve forecasting performance.

Despite these advances, most deep learning models are designed for high-frequency forecasting tasks, such as hourly or daily load prediction, where large volumes of sequential data are available. In contrast, monthly load forecasting typically involves relatively small datasets and heterogeneous tabular predictors, including meteorological, demographic, and economic variables. Under such conditions, deep learning models may be prone to overfitting and often require extensive hyperparameter tuning.

2.3 Tree-Based Ensemble Models for Tabular Load Forecasting

In recent years, tree-based ensemble learning models have demonstrated strong and consistent performance in electricity load forecasting tasks [24]. Compared with linear models, tree-based methods can flexibly capture nonlinear relationships and higher-order interactions without requiring explicit variable transformation or prior specification.

Gradient boosting algorithms such as XGBoost have gained particular attention because of their ability to efficiently model complex tabular datasets containing heterogeneous features. These models are robust to multicollinearity and can effectively handle missing values, making them well suited for operational forecasting environments where data quality and consistency may vary.

Furthermore, tree-based models are particularly advantageous for medium-scale datasets, where the number of observations is limited but the feature space includes diverse explanatory variables. In such scenarios, boosting algorithms often outperform deep learning models by achieving a better balance between predictive accuracy and model stability. Recent studies have also reported that tree-based ensemble models often outperform deep learning architectures on tabular datasets because of their ability to effectively handle heterogeneous features and limited sample sizes [25].

In addition, tree-based models are naturally compatible with explainability tools such as SHAP (SHapley Additive exPlanations), which enable the decomposition of predictions into feature-level contributions. This property makes them especially attractive for applications requiring both predictive performance and interpretability.

2.4 Explainability and Feature Engineering in Load Forecasting

Feature engineering plays a crucial role in improving load forecasting performance. Many studies incorporate lag features, rolling statistics, and weather-related indicators to capture temporal dependencies and seasonal effects in electricity demand [26]. However, in many existing approaches, feature engineering is treated as a one-time preprocessing step rather than an iterative model improvement process.

Recently, explainable machine learning techniques have been increasingly adopted to interpret load forecasting models. SHAP has been widely used to quantify the contribution of input variables to prediction outcomes at both global and local levels [27]. Several studies have applied SHAP to analyze the

influence of meteorological variables, mobility indicators, and economic activity on electricity consumption patterns [28,29]. These studies provide valuable insights into the relative importance of different demand drivers and improve the transparency of machine learning models used in energy forecasting.

However, in most existing work SHAP is primarily used for post-hoc interpretation, meaning that the explainability results are analyzed after model training without directly influencing the model development process. Consequently, although SHAP helps reveal the factors driving model predictions, the insights obtained from these analyses are rarely used to guide systematic feature refinement or targeted model retraining. As a result, explainability and model optimization often remain disconnected steps within the forecasting workflow.

2.5 Research Gap

Despite the growing adoption of machine learning and explainable AI techniques in electricity load forecasting, several limitations remain.

First, many existing studies rely on either deep learning architectures designed for high-frequency forecasting or conventional machine learning models that do not explicitly consider month-specific demand heterogeneity, which can be substantial in seasonal electricity consumption patterns.

Second, feature engineering is often performed as a static preprocessing step, relying on predefined lag variables, statistical transformations, or domain-driven variable selection, without integrating model interpretation into an iterative improvement process.

Third, although SHAP has been widely used for interpreting load forecasting models, its potential role as a diagnostic tool for guiding targeted feature engineering and retraining has been relatively underexplored. In most existing studies, explainability techniques are applied only after the final model is trained, and the insights obtained from the analysis are not systematically fed back into the modeling pipeline to improve predictive performance.

These limitations indicate a lack of a closed feedback loop between model interpretation and model refinement in existing load forecasting research.

To address these limitations, this study proposes a month-conditioned boosting framework integrated with a SHAP-in-the-loop retraining strategy. The proposed approach uses SHAP-based error attribution to identify the most influential features in prediction-vulnerable months and introduces targeted interaction features to refine the model. By embedding explainability directly into the model improvement loop, the framework establishes an iterative process consisting of error diagnosis, explainability analysis, targeted feature construction, and model retraining. This design enhances both predictive performance and interpretability while controlling feature proliferation, providing a data-driven practical framework for monthly electricity load forecasting.

3 Data Description

To evaluate the performance of the proposed framework, we used multidimensional public datasets covering electricity, meteorology, demographics, and economics. All datasets spanned the period from January 2022 to December 2024 and were obtained from officially published public sources to ensure reliability and reproducibility. No missing values were observed in any of the datasets. Detailed data sources and access links are described in the Availability of Data and Materials section.

3.1 Electricity Load Data

Monthly electricity load data (MWh) were obtained from the Korean Public Data Portal, which provides the official national electricity consumption statistics. The dataset contains the aggregated load values for the entire South Korean power system. The original records were available at monthly, daily, and hourly resolutions, and for this study they were aggregated by computing the monthly averages to ensure consistency with the temporal resolution of the other explanatory variables.

3.2 Population Data

Population statistics were obtained from the Ministry of the Interior and the Safety Resident Registration System. The dataset includes residents, individuals with unregistered addresses, and overseas Koreans, but excludes foreign nationals. The original records were reported on a daily basis and were aggregated into monthly averages to align with the temporal resolution of the other explanatory variables.

3.3 Foreign Resident and Visitor Data

Two foreign-related datasets were incorporated. (1) Monthly arrivals and departures of foreign nationals from the Korean Statistical Information Service (KOSIS) and (2) the number of short- and long-term foreign residents from the Korean Public Data Portal.

These indicators reflect the population mobility and international activity levels, which can influence electricity demand. The original records were reported on a daily basis and were aggregated into monthly totals to ensure consistency with the temporal resolution of the other variables.

3.4 Meteorological Data

The meteorological data were obtained from the Open Data Portal of the Korea Meteorological Administration. The dataset originally contained daily records of temperature (°C) and precipitation (mm). For this study, the daily values were aggregated into monthly averages for temperature and monthly totals for precipitation. Station-level measurements were then aggregated at the national level to ensure consistency with the electricity demand data.

3.5 Economic Indicators

Three economic indicators were included in this study: (1) Monthly average exchange rate from the Woori Bank's Foreign Exchange Center; (2) Korea Composite Stock Price Index from Investing.com; and (3) exports and imports from the K-Stat Global Trade Statistics Service.

The exchange rate and KOSPI were originally reported on a daily basis and were aggregated into monthly averages, while the trade statistics were provided on a monthly basis and thus used without further aggregation. These variables were used to capture macroeconomic fluctuations potentially affecting electricity demand.

4 Proposed Framework

As shown in Fig. 1, the workflow of the proposed framework begins with data integration and preprocessing, followed by lag and rolling window feature engineering. A baseline model is then trained and optimized. Error analysis is conducted to identify the months with the highest prediction errors, for which SHAP-guided feature construction is applied by creating interaction variables from the most influential features identified via SHAP value analysis. The model is retrained with the augmented features, and the final

evaluation and interpretation are performed using SHAP value calculations to ensure improved accuracy and interpretability.

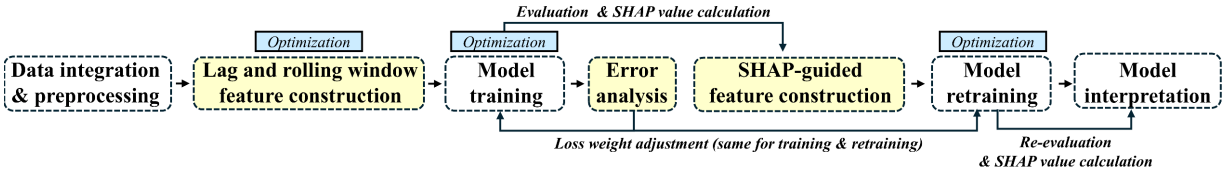

Figure 1: Overall workflow of the proposed month-conditioned boosting framework with SHAP-in-the-loop retraining.

4.1 Data Preprocessing

None of the datasets used in this study contained missing values. Therefore, the imputation procedures were unnecessary. As all variables were already in numeric format, no categorical-to-numeric conversion was required. However, because the datasets were obtained from multiple public platforms, temporal alignment was essential. Each dataset was aggregated or resampled to a monthly resolution to ensure consistency across the variables. Subsequently, the datasets were concatenated into a single integrated data frame and all values were converted to a floating-point format for numerical stability during model training.

All preprocessing operations were implemented in Python using the Pandas library, which was employed for the data manipulation, merging, and type conversion tasks.

4.2 Lag and Rolling Window Feature Construction

To capture the temporal dependencies of electricity demand, we constructed lag and rolling window features from the explanatory variables. Lag features were generated by shifting each variable by a predefined time step, enabling the model to learn from historical patterns. Rolling window statistics, including mean, standard deviation, median, maximum, and minimum, were computed over a fixed period using only past observations to avoid information leakage. To prevent the indiscriminate addition of excessive lag and rolling window features, Optuna was employed to select only those configurations that contributed to improved predictive performance. The optimization process was configured with a lag search range of 1–12 months and a rolling window range of 2–6 months. In each trial, candidate lag values and rolling window lengths were sampled, and leakage-free lag and rolling statistical features were generated accordingly. The model was then trained and evaluated using a rolling-origin time-series cross-validation scheme to ensure temporal consistency.

The optimization objective was defined as minimizing the mean absolute error (MAE) of the out-of-sample predictions. A Tree-structured Parzen Estimator (TPE) sampler with a fixed random seed (42) was used for the search process, and the optimization was conducted for 50 trials, after which the configuration yielding the lowest MAE was selected as the final temporal feature setting. In this study, lag and rolling statistical features were generated for both the explanatory variables and the target variable (monthly electricity load).

4.3 Error Analysis

The primary objective of the error analysis step was to enable loss-weight adjustment for a targeted performance improvement. After initial training, the absolute errors between the predicted and actual values for each month of the evaluation year (2024) were calculated. Months whose errors exceed the overall average error were identified as “vulnerable months” and collected into the set V.

Set V was determined once, based on the baseline evaluation, and remained fixed throughout the retraining process, ensuring that the identification of vulnerable months was not influenced by later model adjustments. The tuning of loss weights for these months was performed using only the training dataset (2022–2023) without referencing the evaluation year (2024), thereby preserving temporal consistency and avoiding data leakage.

For each month in V, an equal additional weight was applied to the loss function during retraining, increasing the emphasis of the model on reducing the prediction errors for these months. The magnitude of the weight adjustment was determined using a grid search. The magnitude of the weight adjustment was determined via a grid search over the candidate set $[\text{None}, 1.5, 2.0, 3.0, 4.0, 5.0]$; the results are presented in the Experiment section.

After retraining, the model was reevaluated on the 2024 dataset, measuring both the overall performance metrics and month-level improvements within V. We used this approach to ensure that the effectiveness of the weighting strategy was validated not only in terms of overall accuracy, but also for specific months in which the model initially exhibited poor performance.

4.4 SHAP-Guided Feature Construction

Following the loss-weight adjustment and model evaluation, SHAP analysis was conducted to identify the dominant factors contributing to the remaining prediction errors. In this study, SHAP-in-the-loop refers to a refinement procedure in which SHAP-based explanation is explicitly incorporated into the model improvement pipeline rather than used solely for post-hoc interpretation. Specifically, SHAP analysis is triggered after the weighted lag/rolling model is evaluated and residual prediction errors are analyzed.

Unlike the previous error analysis stage, which considered all months whose absolute errors exceeded the overall average, the SHAP-guided procedure focuses on the single month exhibiting the largest prediction error in the evaluation period (2024). For this selected month, SHAP values are computed for the corresponding prediction, and the feature with the largest absolute SHAP contribution is identified as the dominant error-related variable. Importantly, SHAP values are computed for predictions generated in the evaluation stage of the rolling-origin validation rather than on the training data, thereby reducing the risk of information leakage in the SHAP-guided feature construction process. In the present implementation, SHAP analysis is performed once per refinement cycle, targeting only the highest-error month to maintain methodological simplicity and interpretability.

An interaction feature is then created between this top-contributing feature and the selected month. This is implemented by multiplying the feature values by a binary indicator representing whether the observation belongs to the target month, thereby capturing the month-specific influence of that feature. Although the implementation can support multiple feature–month interaction pairs, this study adopts a conservative single-pair design, consisting of the highest-error month and its most influential SHAP-identified feature. This choice was made to preserve interpretability and to avoid excessive feature proliferation, which may increase model complexity and overfitting risk in a relatively small monthly dataset. In this study, the feature identified by SHAP for the highest-error month was Total Electricity Monthly Average (lag12), which exhibited the largest absolute SHAP contribution to the prediction error. The interaction feature was therefore constructed using this variable and the corresponding month indicator. This targeted feature construction strategy prioritizes the most influential error-related factor while maintaining a parsimonious and interpretable model structure.

The newly created interaction feature is appended to the training set, and the model is retrained using the same loss-weight adjustment scheme used in the initial training phase. This procedure enables the

model to explicitly capture the interaction between a critical feature and the month most prone to large prediction errors, thereby reducing targeted prediction bias while maintaining temporal consistency in the training process. The overall SHAP-in-the-loop refinement procedure, including the trigger condition, feature selection mechanism, and retraining step, is summarized in Algorithm 1.

Algorithm 1: SHAP-in-the-loop month-conditioned refinement

Require: Monthly dataset $D = \{(x_t, y_t, m_t)\}_{t=1}^{T=1}$
Ensure: Final forecasting model $f^\star$
1: Generate lag and rolling features using past observations only
2: Train baseline XGBoost model and compute month-wise absolute errors
3: Identify vulnerable months with above-average errors
4: Apply higher loss weights to vulnerable months and retrain the model
5: Select the month $m^\star$ with the largest residual error
6: Compute SHAP values for the prediction corresponding to $m^\star$
7: Select the feature $j^\star$ with the largest absolute SHAP contribution
8: Create interaction feature $z_t = x_{t,,j^\star} \cdot 1(m_t = m^\star)$
9: Retrain the model using the augmented feature set
10: Evaluate the final model
11: return $f^\star$

4.5 Regression Model

In this study, a prediction model is implemented using Extreme Gradient Boosting (XGBoost) regression, which is a tree-based ensemble method based on gradient boosting. Given a dataset $\{(x_i, y_i)\}_{i=1}^{N}$, the model aims to find an additive function $F_M(x)$ comprising M regression trees $f_m \in \mathcal{F}$, where $\mathcal{F}$ denotes the space of regression trees as shown in Eq. (1):

$$F_M(\boldsymbol{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}), f_m \in \mathcal{F}, \tag{1}$$

The optimization objective combines the training loss $l(y_i, \widehat{y_i})$ and a regularization term $\Omega(f_m)$ to control model complexity as shown in Eq. (2):

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \widehat{y_i}) + \sum_{m=1}^{M} \Omega(f_m), \Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2, \tag{2}$$

where T is the number of leaves, w is the leaf weight vector, and γ and λ are regularization parameters.

Both the initial model training after error analysis and the retraining after SHAP-guided feature construction employed Optuna for hyperparameter optimization. The search space included 300–900 estimators, learning rates from 0.01 to 0.3, maximum tree depths of 3–8, minimum child weights of 1.0–10.0, subsample and column-sampling ratios between 0.6 and 1.0, L1 and L2 regularization terms α and λ from 0.0 to 1.0 and 0.5 to 3.0, respectively, and γ values between 0.0 and 5.0. The optimization results are presented in detail in the Experiment section.

4.6 Model Training and Evaluation Methodology

We evaluate all the stages using a one-step-ahead rolling-origin procedure with an expanding training window. Let t indicate the index months. Starting with a warm-up period of 12 months, at each step t the

model is trained on $\{1, \ldots, t-1\}$ and predicts y_t. This generates an out-of-sample sequence of month-ahead predictions without look-ahead. Therefore, the protocol is leakage-free and aligned with how the rolling lag and rolling-window features are constructed from past observations only.

To focus learning on prediction-vulnerable months when needed, we allow timestamp-dependent sample weights during fitting. If a weighting function is supplied, the per-sample weights are computed from the training indices and passed to the learner; otherwise, uniform weights are used.

The performance is reported for the concatenated out-of-sample predictions using the root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (), and mean absolute percentage error (MAPE). RMSE and MAE are expressed in the original units of electricity demand (MWh), whereas MAPE is reported as a percentage.

We also compute monthly summaries to analyze where improvements occur across the three stages (baseline, lag and rolling feature construction, and SHAP-guided interaction construction).

5 Experiment and Discussion

This section presents the experimental setup, including the optimized feature augmentation and loss weight adjustment, as well as the optimized hyperparameters for both model training and retraining. We then provide a performance evaluation and detailed analysis of the proposed model.

5.1 Experimental Setup

Table 1 lists the optimization results of the feature augmentation criteria, month-specific loss weights, and hyperparameters used in this study.

Table 1: Optimization results for feature augmentation criteria and model hyperparameters.

Category	Parameter Optimal Value	
Feature engineering	Lag length: 12 Rolling window size: 3 SHAP-guided Interaction feature: Total Electricity monthly average (lag12) × Months 8	
Error analysis & weight adjustment	Target Months: 5, 8, 9, 11 Loss weight applied to target months: 4 (others = 1)	
XGBoost hyperparameters	**Initial Training**	**Retraining**
	n_estimators: 400 learning_rate: 0.2 max_depth: 6 subsample: 0.7 colsample_bytree: 1 min_child_weight: 10 reg_alpha: 0.1 reg_lambda: 1 gamma: 3	n_estimators: 300 learning_rate: 0.3 max_depth: 6 subsample: 0.7 min_child_weight: 10 colsample_bytree: 0.7 reg_alpha: 0.3 reg_lambda: 1 gamma: 1

The three key components of feature augmentation are as follows:

1. Lag length: The number of months of historical targets and explanatory variables were used as lag features. A 12-month lag was selected to capture the long-term seasonal dependencies in electricity demand.
2. Rolling window size: The size of the temporal averaging window was applied to lagged variables to smooth short-term fluctuations while retaining trend information. The optimal tumor size was determined to be three months.
3. SHAP-guided interaction feature: A month-conditioned interaction term was constructed by multiplying the most influential feature (identified via SHAP analysis) by a binary indicator for the month with the highest prediction error. Targeted feature engineering was designed to enhance the sensitivity of the model to specific seasonal effects.

For loss weight adjustment, target months for which absolute prediction errors exceeded the overall average error in the baseline evaluation were assigned higher weights during training to prioritize error reduction. Specifically, May, August, September, and November were weighed at 4, whereas all other months were weighed at 1. The optimal hyperparameters for both the initial training and retraining were obtained using Optuna.

5.2 Baseline Model Performance

To validate the suitability of XGBoost for the integrated dataset used in this study, we conducted a baseline performance comparison between XGBoost and a set of representative models, including both linear regressors and other tree-based methods. The compared models included Linear Regression, Ridge, Lasso, Random Forest, Support Vector Regression (SVR), LightGBM, and CatBoost. All models were trained using the rolling-origin cross-validation scheme, and default hyperparameters were applied for a fair comparison.

The results listed in Table 2 demonstrate that XGBoost achieves competitive accuracy among tree-based models. Among all the evaluated models, linear approaches (Linear Regression, Ridge, Lasso) perform poorly, exhibiting high error values and negative R^2 scores. The tree-based models generally deliver superior results, with XGBoost achieving the best balance of RMSE, MAE, and R^2, thereby justifying its selection as the primary model for subsequent optimization and enhancement steps.

Table 2: Baseline regression performance on the integrated dataset (pre-optimization).

Model	RMSE ($\times 10^9$)	MAE	R^2	MAPE (%)
Linear regression	178.26	216,566.71	−10.138	13.431
Ridge	177.16	215,475.32	−10.069	13.362
Lasso	39.75	147,734.49	−1.483	9.354
Random forest	8.07	70,770.02	0.496	4.623
XGBoost	7.13	70,133.78	0.555	4.579
SVR	17.46	114,693.04	−0.091	7.386
LightGBM	16.92	113,658.38	−0.057	7.438
CatBoost	10.06	80,938.10	0.372	5.252

It should be noted that the baseline models were evaluated using standard default configurations to provide a transparent comparison across different model families. Tree-based models such as Random Forest, XGBoost, LightGBM, and CatBoost are inherently robust to feature scaling and nonlinear relationships,

whereas purely linear models may be less suitable for capturing the complex interactions and seasonal effects present in electricity demand data. Consequently, the negative R^2 values observed for several linear models mainly reflect the strong nonlinear characteristics of the forecasting problem rather than deficiencies in preprocessing. Since the primary objective of this study is to evaluate the proposed SHAP-in-the-loop refinement framework within a representative forecasting model, XGBoost was selected based on its stable performance across the baseline comparisons.

5.3 Effect of Lag and Rolling Window Features

In this experiment, the XGBoost model is trained using the optimized hyperparameters listed in Table 1. Table 3 compares the baseline model (with optimized hyperparameters only) with the models incorporating lag and rolling window features, along with the loss weight adjustment derived from the error analysis.

Table 3: Performance comparison of XGBoost baseline and enhanced models.

Model	RMSE ($\times 10^9$)	MAE	R^2	MAPE (%)
XGBoost (Baseline, tuned)	6.45	62,973.95	0.64	3.97
With Lag & rolling Features	2.50	44,479.18	0.86	2.86

As listed in Table 3, the optimized XGBoost baseline model achieves a better performance compared with the default baseline listed in Table 2. However, after incorporating lag features, rolling window features, and loss weight adjustment, the model performance improves substantially across all evaluation metrics. The results demonstrate that adding lag and rolling window features, together with error-analysis-based weight adjustment, significantly enhances the predictive performance of the model.

5.4 Effect of the SHAP-Guided Interaction Feature

To realize the final goal of the SHAP-in-the-loop framework, we incorporated a SHAP-guided interaction feature into the model. Among the vulnerable months identified in the previous stage, August exhibited the highest residual prediction error after applying lag and rolling features together with the loss-weight adjustment scheme. Therefore, August was selected as the target month for SHAP-guided refinement.

As illustrated in Fig. 2, SHAP analysis revealed that Total Electricity Monthly Average (lag12) had the largest contribution to the prediction error for this month. Based on this result, a month-conditioned interaction feature defined as Total Electricity Monthly Average (lag12) × August was generated and incorporated into the model. The model was then retrained using the optimized hyperparameters.

As listed in Table 4, the performance improves further compared with that of the previous step. Although the gain is moderate, the results confirm that the SHAP-guided feature construction contributes to additional error reduction and supports the intended in-the-loop process of the proposed method.

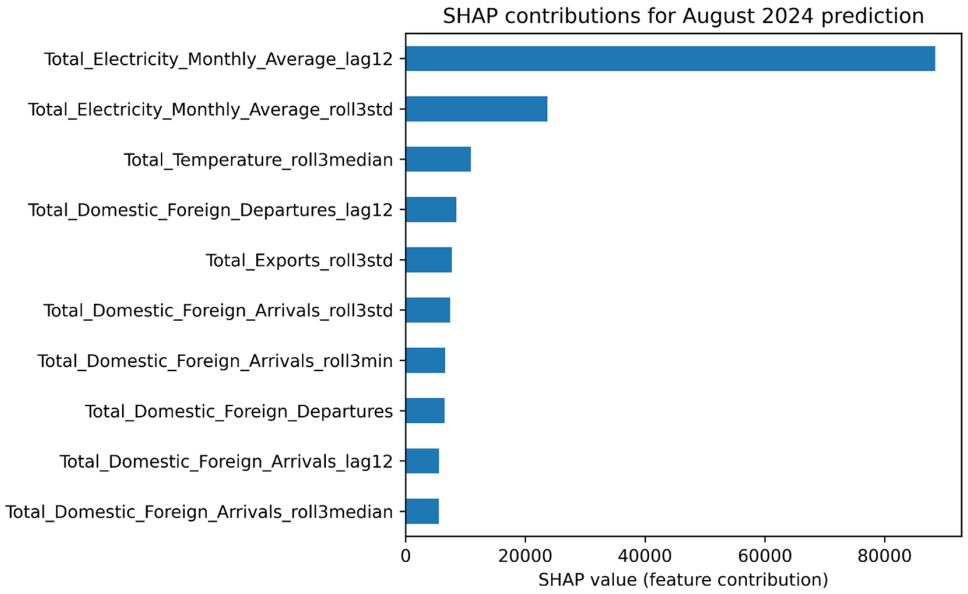

Figure 2: Local SHAP explanation for the electricity load prediction in August 2024. The plot illustrates the contribution of the most influential features to the prediction for the selected observation. Positive SHAP values indicate factors increasing the predicted load, while negative values indicate factors decreasing it. SHAP values are expressed in the same unit as the prediction target (electricity load, MWh).

Table 4: Model performance with SHAP-guided interaction feature.

Model	RMSE ($\times 10^9$)	MAE	R^2	MAPE (%)
With SHAP-guided interaction feature	2.38	43,748	0.87	2.85

5.5 Comparative Analysis across Model Enhancement Steps

Fig. 3 shows the monthly absolute error profiles across the three model enhancement steps: the baseline XGBoost model, the model with lag and rolling features, and the retrained model with SHAP-guided interaction features. As shown in Fig. 3, the baseline model shows high variability with pronounced error peaks, particularly in May, August, September, and November. By incorporating the lag and rolling features, the error distribution becomes more stable, with overall reductions in both the mean and variance. Finally, retraining with SHAP-guided interaction features leads to further error reduction, as reflected by the lower average absolute error (green line), particularly in the months in which the baseline model previously exhibits the highest errors.

To assess the reliability of the forecasting improvements, we analyzed the distribution of errors obtained from the rolling-origin evaluation. Because this evaluation scheme generates multiple out-of-sample forecasts across different time points, it enables the estimation of confidence intervals for the forecasting metrics.

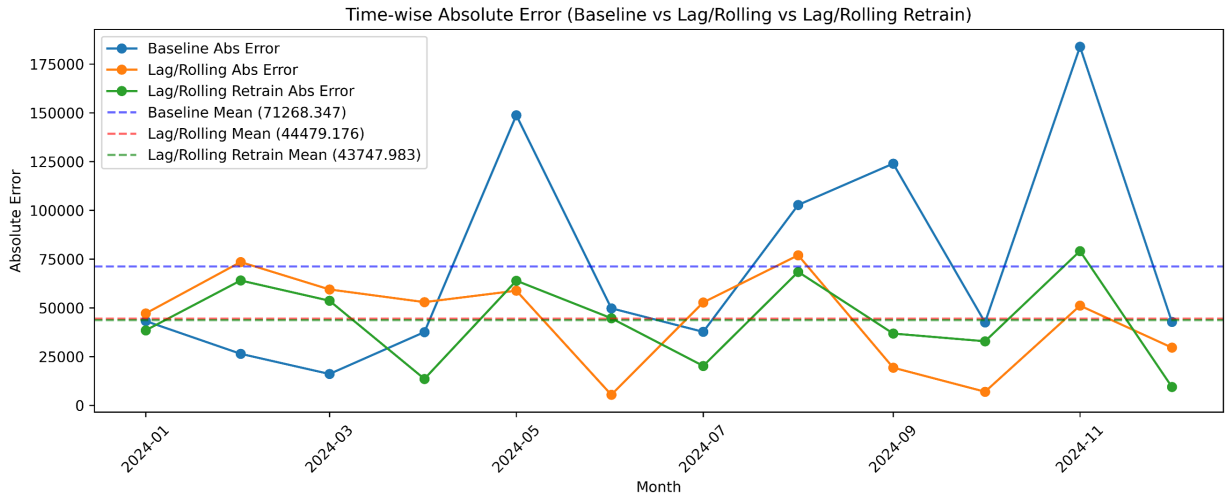

Figure 3: Monthly absolute error comparison of baseline, lag/rolling, and retrained models.

Table 5 presents the forecasting performance of the baseline model and the proposed lag/SHAP-guided model together with 95% confidence intervals estimated from the rolling-origin evaluation errors. The results indicate that the proposed model consistently outperforms the baseline across all evaluation metrics. In particular, the RMSE decreased from 88,320 (95% CI: 50,167–118,433) to 50,036 (95% CI: 38,809–59,743), while the MAE decreased from 71,268 (36,650–105,887) to 44,479 (29,271–59,687). Similarly, the MAPE improved from 4.69% (2.28–7.09) to 2.86% (1.89–3.84).

Table 5: Forecasting performance comparison with confidence intervals and relative improvement.

Model	RMSE (95% CI)	MAE (95% CI)	MAPE (%) (95% CI)
Baseline	88,320 (50,167–118,433)	71,268 (36,650–105,887)	4.69 (2.28–7.09)
Lag/SHAP-guided model	50,036 (38,809–59,743)	44,479 (29,271–59,687)	2.86 (1.89–3.84)
Improvement (%)	67.9	37.6	38.9

To facilitate practical interpretation, the relative improvements are also reported in Table 5. Compared with the baseline configuration, the proposed model achieved a 67.9% reduction in RMSE, a 37.6% reduction in MAE, and a 38.9% reduction in MAPE, indicating substantial forecasting performance gains.

To further assess the reliability of these improvements, paired statistical tests were conducted on the rolling-origin forecast errors across the common evaluation months. Although the proposed model consistently produced lower errors in most forecasting windows, the statistical tests did not reach conventional significance levels ($p > 0.05$), likely due to the relatively small number of rolling-origin evaluation points ($n = 12$). Nevertheless, the consistent reduction in forecast errors together with narrower confidence intervals suggests that the proposed SHAP-guided feature refinement contributes to improved forecasting stability.

It should also be noted that electricity demand in this study is measured in system-scale units (MWh) for the entire national grid. Consequently, the numerical magnitude of the RMSE values appears large but corresponds to realistic deviations in nationwide electricity demand forecasting.

5.6 Model Interpretability

Fig. 4 shows the monthly variation in SHAP importance for the top 10 features extracted from the retrained model. The results shown in Fig. 4 reveal distinct seasonal patterns: mobility-related indicators, such as Total Domestic and Foreign Arrivals (rolling 3-month mean and max), exhibit high importance, particularly during the summer and winter months, which are periods of peak energy demand. This suggests that large-scale domestic and international travel, which is closely associated with tourism and holiday seasons, substantially influence electricity consumption patterns.

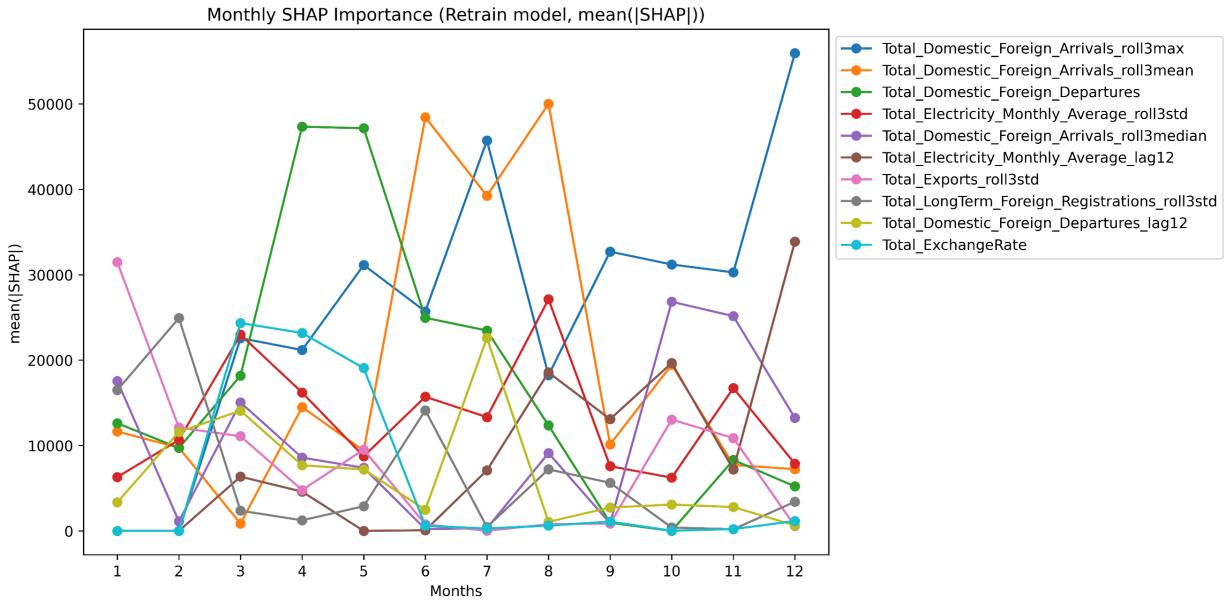

Figure 4: Global SHAP importance of the top 10 features in the retrained model. Feature importance is measured as the mean absolute SHAP value (mean(|SHAP|)) aggregated across the evaluation samples. Larger values indicate a stronger overall contribution of the feature to the model predictions. SHAP values are expressed in the same unit as the target variable (electricity load, MWh).

In contrast, Total Electricity Monthly Average (lag12) consistently contributes across months, reflecting the strong role of historical seasonal demand in shaping future consumption trends. The lag-12 feature effectively captures annual cyclicality, indicating that the consumption of past year remains a robust predictor of current demand.

These findings show that the model learns not only from past electricity use but also from mobility-driven changes, effectively capturing the interactions of seasonal demand cycles with broader socio-economic activity. This improves the interpretability and highlights the practical importance of considering both mobility and seasonality in electricity demand forecasting.

5.7 Limitations

Despite the promising results, this study has several limitations. First, the dataset covers only 36 months, which represents a relatively short observation period for monthly electricity load forecasting. A longer historical record could potentially capture additional structural demand changes and seasonal variability.

Second, although the proposed framework mitigates overfitting risks by introducing only a single SHAP-guided interaction feature, the results may still be influenced by the limited data size. To assess model stability, we employed a rolling-origin time-series validation scheme that repeatedly trains and evaluates the model

using temporally ordered splits. This approach allows the model performance to be examined across multiple forecasting windows while preserving temporal causality.

Third, the current implementation adopts a conservative single-pair SHAP-guided interaction strategy. While the framework can be extended to construct multiple feature–month interaction terms, the present study intentionally restricts the interaction construction to the most influential SHAP-identified feature in the highest-error month. This design choice was made to maintain interpretability and to avoid excessive feature proliferation in a relatively small dataset.

Finally, because SHAP explanations are inherently model-dependent, the interaction feature identified in this study should be interpreted as a model-guided refinement rather than a definitive causal relationship between variables.

Future studies may address these limitations by exploring richer interaction structures, such as multiple SHAP-guided feature–month interactions, higher-order interaction terms, or model-driven interaction discovery methods. In addition, systematic sensitivity or ablation analyses could be conducted to evaluate the robustness of different feature-selection strategies. Furthermore, validating the proposed SHAP-in-the-loop framework using longer historical time series or multi-regional datasets would help assess its generalizability and robustness under varying data conditions.

6 Conclusion

This study proposes an explainability-in-the-loop framework for short-term electricity demand forecasting by integrating error analysis, feature engineering, and SHAP-based interpretation into the model development pipeline. Starting from a baseline XGBoost model, we demonstrated that the inclusion of lag and rolling window features, coupled with targeted loss weight adjustments for vulnerable months, substantially improved predictive performance. Further enhancement was achieved by constructing SHAP-guided interaction features, which helped the model capture the critical relationships between historical consumption and external drivers of demand.

The experimental results confirmed that each stage of model refinement contributed to reduced error and increased robustness, with the most significant improvements observed after combining temporal features and error-driven weighting. Beyond predictive accuracy, the SHAP analysis provided valuable interpretability by revealing the month-specific importance of mobility-related and seasonal features, thereby offering insights into how electricity demand is shaped by both structural and socio-economic dynamics.

Overall, the findings highlight that explainable-feature-driven optimization can enhance both the accuracy and transparency of electricity demand forecasting models. This approach strengthens the reliability of forecasts in practice and facilitates a deeper understanding of the underlying factors driving demand variability. In conclusion, the proposed framework offers a data-driven solution for electricity demand forecasting in practical settings. Future work may extend this framework to multi-regional datasets, incorporate real-time mobility and climate indicators, and explore adaptive weighting strategies for further robustness.

Acknowledgement: The authors gratefully acknowledge the support of Hanyang University, Seoul, Republic of Korea.

Funding Statement: This research was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2023-00275579), and in part by the research fund of Hanyang University (HY-202500000003761).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jinsung Park, Jaehyuk Lee and Eunchan Kim; methodology, Jinsung Park, Jaehyuk Lee and Eunchan Kim; software,

Jinsung Park; validation, Jinsung Park and Jaehyuk Lee; formal analysis, Jinsung Park and Jaehyuk Lee; investigation, Jinsung Park and Jaehyuk Lee; resources, Eunchan Kim; data curation, Jinsung Park; writing—original draft preparation, Jinsung Park and Jaehyuk Lee; writing—review and editing, Jinsung Park, Jaehyuk Lee and Eunchan Kim; visualization, Jinsung Park; supervision, Eunchan Kim; project administration, Eunchan Kim; funding acquisition, Eunchan Kim. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in public repositories. Monthly electricity load data are available from the Korea Public Data Portal (https://www.data.go.kr). Population statistics were obtained from the Ministry of the Interior and Safety (https://jumin.mois.go.kr). Foreign nationals statistics, including monthly arrivals and departures, were retrieved from Statistics Korea (https://kosis.kr), and data on short- and long-term foreign residents are available from the Korea Public Data Portal (https://www.data.go.kr). Meteorological data, including monthly average temperature and accumulated precipitation, were obtained from the Korea Meteorological Administration (https://data.kma.go.kr). Economic indicators, including monthly average exchange rates, the Korea Composite Stock Price Index, and monthly export and import totals, were retrieved from publicly accessible financial and trade statistics platforms.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Berk K. Modeling and forecasting electricity demand: a risk management perspective. Cham, Switzerland: Springer; 2015. doi:10.1007/978-3-658-08669-5.
2. Hong T, Dickey DA. Electric load forecasting: fundamentals and best practices. Cary, NC, USA: SAS Institute Inc.; 2012.
3. Kim Y, Choi Y, Min SK. Future changes in heat wave characteristics and their impacts on the electricity demand in South Korea. Weather Clim Extrem. 2022;37(8):100485. doi:10.1016/j.wace.2022.100485.
4. Hiruta Y, Ishizaki NN, Ashina S, Takahashi K. Regional and temporal variations in the impacts of future climate change on Japanese electricity demand: simultaneous interactions among multiple factors considered. Energy Convers Manag X. 2022;14(1):100172. doi:10.1016/j.ecmx.2021.100172.
5. Chang Y, Kim CS, Miller JI, Park JY, Park S. A new approach to modeling the effects of temperature fluctuations on monthly electricity demand. Energy Econ. 2016;60:206–16. doi:10.1016/j.eneco.2016.09.016.
6. Cawthorne D, de Queiroz AR, Eshraghi H, Sankarasubramanian A, DeCarolis JF. The role of temperature variability on seasonal electricity demand in the southern US. Front Sustain Cities. 2021;3:644789. doi:10.3389/frsc.2021.644789.
7. Rana PS, Kalpana Chahat, Modi SK, Yadav AL, Singla S. Comparative analysis of tree-based models and deep learning architectures for tabular data: performance disparities and underlying factors. In: Proceedings of the 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech); 2023 Dec 23–24; Banur, India. p. 224–31. doi:10.1109/ICACCTech61146.2023.00044.
8. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 785–94. doi:10.1145/2939672.2939785.
9. Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms. Appl Energy. 2018;221(C):386–405. doi:10.1016/j.apenergy.2018.06.131.
10. Yaprakdal F, Arısoy MV. A multivariate time series analysis of electrical load forecasting based on a hybrid feature selection approach and explainable deep learning. Appl Sci. 2023;13(23):12964. doi:10.3390/app132312946.
11. Zabin R, Haque KF, Abdelgawad A. PredXGBR: a machine learning framework for short-term electrical load prediction. Electronics. 2024;13(22):4521. doi:10.3390/electronics13224521.

12. Moon J, Rho S, Baik SW. Toward explainable electrical load forecasting of buildings: a comparative study of tree-based ensemble methods with Shapley values. Sustain Energy Technol Assess. 2022;54(3):102888. doi:10.1016/j.seta.2022.102888.
13. Akhtar S, Shahzad S, Zaheer A, Ullah HS, Kilic H, Gono R, et al. Short-term load forecasting models: a review of challenges, progress, and the road ahead. Energies. 2023;16(10):4060. doi:10.3390/en16104060.
14. Ziel F, Liu B. Lasso estimation for GEFCom2014 probabilistic electric load forecasting. Int J Forecast. 2016;32(3):1029–37. doi:10.1016/j.ijforecast.2016.01.001.
15. Alfares HK, Nazeeruddin M. Electric load forecasting: literature survey and classification of methods. Int J Syst Sci. 2002;33(1):23–34. doi:10.1080/00207720110067421.
16. Yang A, Li W, Yang X. Short-term electricity load forecasting based on feature selection and Least Squares Support Vector Machines. Knowl Based Syst. 2019;163(11):159–73. doi:10.1016/j.knosys.2018.08.027.
17. Chen K, Chen K, Wang Q, He Z, Hu J, He J. Short-term load forecasting with deep residual networks. IEEE Trans Smart Grid. 2018;10(4):3943–52. doi:10.1109/tsg.2018.2844307.
18. Smyl S, Dudek G, Pełka P. ES-dRNN: a hybrid exponential smoothing and dilated recurrent neural network model for short-term load forecasting. IEEE Trans Neural Netw Learning Syst. 2023;35(8):11346–58. doi:10.1109/tnnls.2023.3259149.
19. Waheed W, Xu Q, Aurangzeb M, Iqbal S, Dar SH, Elbarbary ZMS. Empowering data-driven load forecasting by leveraging long short-term memory recurrent neural networks. Heliyon. 2024;10(24):e40934. doi:10.1016/j.heliyon.2024.e40934.
20. Bouktif S, Fiaz A, Ouni A, Serhani MA. Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: comparison with machine learning approaches. Energies. 2018;11(7):1636. doi:10.3390/en11071636.
21. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271. 2018. doi:10.48550/arxiv.1803.01271.
22. Wan A, Chang Q, AL-Bukhaiti K, He J. Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism. Energy. 2023;282:128274. doi:10.1016/j.energy.2023.128274.
23. Chen J, Liu L, Guo K, Liu S, He D. Short-term electricity load forecasting based on improved data decomposition and hybrid deep-learning models. Appl Sci. 2024;14(14):5966. doi:10.3390/app14145966.
24. Muqtadir A, Bin L, Zhou Y, Chen S, Kazmi SN. Nowcasting the next hour of residential load using boosting ensemble machines. Sci Rep. 2025;15(1):7157. doi:10.1038/s41598-025-91767-6.
25. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? Adv Neural Inf Process Syst. 2022;35:507–20. doi:10.52202/068431-0037.
26. Nti IK, Teimeh M, Nyarko-Boateng O, Adekoya AF. Electricity load forecasting: a systematic review. J Electr Syst Inf Technol. 2020;7(1):13. doi:10.1186/s43067-020-00021-8.
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
28. Bhupatiraju A, Ahn SB. Explainability-driven feature engineering for mid-term electricity load forecasting in ERCOT's SCENT region. arXiv:2507.22220. 2025. doi:10.48550/arxiv.2507.22220.
29. Alba EL, Oliveira GA, Ribeiro MHDM, Rodrigues ÉO. Electvricity consumption forecasting: an approach using cooperative ensemble learning with SHapley additive exPlanations. Forecasting. 2024;6(3):839–63. doi:10.3390/forecast6030042.