sustainability

MDPI

Article

Prediction of Transformer Load-Rate and Analysis of Meteorological Feature Contributions for Sustainable Operation of Regional Substations: An Explainable Machine Learning Approach

Zhongmin Chen [1], Lei Li [1], Yixian Wang [2], Yunfei Ding [2,*] and Huijun Wu [2]

[1] Maoming Power Supply Bureau of Guangdong Grid Co., Ltd., Maoming 525000, China; 13727802693@139.com (Z.C.); 13580020513@139.com (L.L.)
[2] School of Civil Engineering and Transportation, Guangzhou University, Guangzhou 510006, China; 1112516087@e.gzhu.edu.cn (Y.W.); hjwu@gzhu.edu.cn (H.W.)
* Correspondence: dingyf@gzhu.edu.cn; Tel.: +86-13342884050

Abstract

Prediction of transformer load-rate at regional substations can provide information to support operating condition assessment, capacity margin analysis, and the identification of potential high-load risks. This study selected a 40 MVA main transformer at the Panzhou regional substation in Maoming City, Guangdong Province, China, as the research case. Based on hourly transformer load-rate and observed meteorological data from 2024, the forecasting performance of various time-series baselines and machine learning models was compared. The Random Forest (RF) model was then selected for input feature ablation experiments and explainability analysis. On this basis, Shapley Additive Explanations (SHAP) and Accumulated Local Effects (ALE) were employed to analyze the contributions of meteorological features to model predictions and their nonlinear response relationships. The model used observed meteorological data at the target time as inputs; therefore, the reported results reflect load-rate forecasting performance conditional on observed meteorological conditions. The results show that the RF model achieved a coefficient of determination (R^2) of 0.9654, a mean absolute percentage error (MAPE) of 3.93%, a mean absolute error (MAE) of 0.97 percentage points, and a root mean square error (RMSE) of 1.46 percentage points on the independent test set for 2024. On the independent validation set from June to August 2025, the R^2 and MAPE were 0.9296 and 3.72%, respectively, indicating that the model retained a certain degree of predictive capability on warm-season data from a different year at the same site. The ablation experiments showed that historical load features were the primary source of the model's predictive performance, with an R^2 of 0.9510 obtained using all historical load features alone. After incorporating temporal and meteorological features, the R^2 further increased to 0.9654. SHAP analysis indicated that temperature and relative humidity had relatively high importance among the meteorological variables, while ALE analysis showed that the model exhibited different nonlinear responses to individual meteorological variables. The findings may provide data-driven support for reliable and sustainable operation and load management of regional substations.

Keywords: regional substation transformers; load-rate of transformers; meteorological parameters; random forest regression model; interpretable machine learning; sustainable operation

Check for updates

Academic Editors: Giouli Mihalakakou and Babatunde Olubayo

Received: 7 September 2026
Revised: 21 October 2026
Accepted: 29 October 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Electricity is a major form of final energy consumption [1]. As key nodes connecting the upstream power grid with end users in distribution networks, regional substations are responsible for distributing electricity to residential, commercial, and industrial users within their service areas. Transformer load-rate reflects equipment capacity utilization, operating margin, and potential overload risk, and is therefore an important indicator for evaluating substation operating conditions and power supply reliability. Transformer load-rate forecasting at regional substations can provide information on equipment loading levels and their variation trends, serving as a reference for assessing transformer operating conditions, analyzing capacity margins, and identifying potential overload risks. It can also provide supplementary information for regional distribution network management and related operational dispatch decisions [2,3].

Regional substations supply electricity to different types of users within a given area. Regional end-use electricity demand is associated with various factors, including weather conditions, holidays, and the level of economic activity, and these factors are also related to variations in substation transformer load-rate [4]. In general, regional substation loads can be classified into year-round loads and seasonal loads. The former are associated with daily activity schedules, including electricity consumption by lighting, office equipment, elevators, and industrial production equipment, whereas the latter are closely related to meteorological conditions, such as electricity consumption by heating, ventilation, and air-conditioning systems. Previous studies have shown that electricity demand exhibits typical time-series characteristics and is jointly associated with seasonal effects, weather conditions, temporal patterns, and differences in user behavior. Therefore, historical loads, meteorological parameters, and temporal features should be considered comprehensively in load forecasting [5]. Hourly transformer load-rate typically exhibits time-series characteristics, including trends, periodic patterns, and local fluctuations. Its predictive information is primarily derived from the temporal continuity and periodic structures of historical load-rate data, together with supplementary information provided by external variables, such as temporal and meteorological features.

In recent years, changes in end-use electricity demand have received increasing attention with the adjustment of energy consumption structures and the digitalization of power systems. According to the International Energy Agency (IEA), global electricity demand increased by 4.3% year-on-year in 2024 [6]. Meanwhile, electricity demand varies across regions in association with differences in industrial structures, user types, and climatic conditions, exhibiting pronounced spatial and temporal variations [7–9]. Among these factors, meteorological conditions are closely associated with electricity demand for building cooling and heating. Previous studies have employed multiscale regression and other methods to examine the relationships between climatic conditions and electricity demand in residential and commercial buildings, identifying significant associations between temperature variations and related energy demand [10,11]. In addition, previous research has suggested that temperature variations may be associated with changes in long-term electricity demand levels and consumption patterns [12]. These studies provide a reference for understanding the relationships between meteorological conditions and end-use electricity demand. However, electricity consumption characteristics differ across load types and operational levels, indicating the need for further analysis of transformer load-rate at regional substations.

With the development of machine learning techniques, data-driven methods have been widely applied to electricity load forecasting, building energy consumption prediction, and load forecasting for integrated energy systems. Based on model structures and feature-processing approaches, existing methods can be broadly classified into traditional

machine learning, deep learning, and hybrid forecasting methods. Traditional machine learning methods, represented by Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), and Random Forest (RF), can process multidimensional input features and characterize nonlinear relationships between input variables and prediction targets [13,14]. Deep learning methods, including Long Short-Term Memory (LSTM) networks and attention-based models, primarily focus on extracting temporal dependencies from time-series data [15,16]. In addition, some studies have developed hybrid forecasting frameworks by combining time-series decomposition, feature selection, and different models to improve predictive performance under complex load conditions. Different types of models have distinct characteristics in terms of feature processing, data requirements, and computational complexity. Specifically, sequence models such as LSTM are suitable for learning temporal dependencies, while tree-based models such as XGBoost and RF can effectively process structured input features. Hybrid models generally require more complex model configurations and parameter tuning. Therefore, model selection should consider dataset size, feature types, prediction tasks, and model interpretability requirements.

In practical applications, existing studies have investigated various forecasting scenarios, including building loads, power-grid loads, integrated energy system loads, and microgrid loads. For example, Guo et al. [17] applied a multivariable method to office building load prediction, while Wang et al. [18] employed an ensemble model for commercial building energy consumption prediction. Wu et al. [19] investigated elec-tricity consumption forecasting for campus building clusters using a combined fore-casting model. Related research has addressed coordinated optimization of buildings and integrated community energy systems [20], fault-risk prediction for distribution networks under persistent high-temperature conditions [21], and power-equipment state recognition using transfer learning [22]. In power system load forecasting, hybrid deep learning approaches have also been investigated for grid-connected mul-ti-microgrids [23]. These studies illustrate applications of data-driven approaches to energy-system forecasting, operational optimization, and equipment-state assessment. The forecasting studies provide methodological references for combining historical load, temporal, and external environmental information in load prediction. As model complexity increases, greater attention has also been paid to the basis of model predic-tions and the roles of input variables. In recent years, explainable machine learning methods, such as Shapley Additive Explanations (SHAP), have been increasingly ap-plied to load forecasting studies to quantify the contributions of input features to mod-el outputs [24,25].

Overall, existing load forecasting studies have mainly focused on system-level electricity loads, building energy consumption, household electricity demand, and loads in integrated energy systems. These studies have primarily emphasized improvements in forecasting accuracy, extraction of temporal features, and optimization of model structures. For transformer load-rate forecasting at regional substations, historical load information can characterize the temporal continuity and periodicity of load variations, while the incremental predictive information provided by meteorological variables and their associations with historical load features warrant further investigation. In addition, the relative contributions of different meteorological variables to model predictions and their nonlinear response relationships may be jointly associated with feature correlations and data distributions, requiring a comprehensive evaluation using feature ablation experiments and model interpretability methods. Therefore, a systematic analysis of the predictive contributions of historical load, temporal, and meteorological features from the perspectives of forecasting performance and model interpretability is necessary to provide a basis for transformer load-rate forecasting at regional substations and the appropriate utilization of meteorological information.

Based on these considerations, this study selected the Panzhou regional substation in Maoming City, Guangdong Province, China, as a case study. Using hourly transformer load-rate data and corresponding meteorological observations from 2024, a Random Forest (RF) forecasting model integrating historical load, temporal, and meteorological features was developed. Multidimensional performance evaluation and explainability analysis were conducted to examine the predictive contributions of different input features and the nonlinear response relationships associated with meteorological variables. First, model performance under different temporal and operating conditions was evaluated through benchmark model comparisons, rolling-origin validation, cross-year validation, and evaluations under different operating conditions. Second, input feature ablation experiments were conducted to quantitatively assess the relative contributions and incremental predictive value of historical load, temporal, and meteorological information to model performance. Finally, Shapley Additive Explanations (SHAP) and Accumulated Local Effects (ALE) were employed to analyze the attribution contributions of meteorological variables to model predictions and their nonlinear associations with model outputs. This study conducted hourly load-rate forecasting conditional on observed meteorological conditions at the target time, providing case-based evidence for transformer load-rate forecasting and meteorological feature interpretation at regional substations.

The remainder of this paper is organized as follows. Section 1 presents the introduction. Section 2 describes the data sources, feature engineering, RF regression model, ablation experiments, and explainability analysis methods. Section 3 evaluates the model's load-rate forecasting performance, analyzes the ablation results for different input feature combinations, and discusses the contributions of meteorological parameters to model predictions and their response characteristics using SHAP and ALE. Section 4 presents the conclusions. The overall research framework is shown in Figure 1.

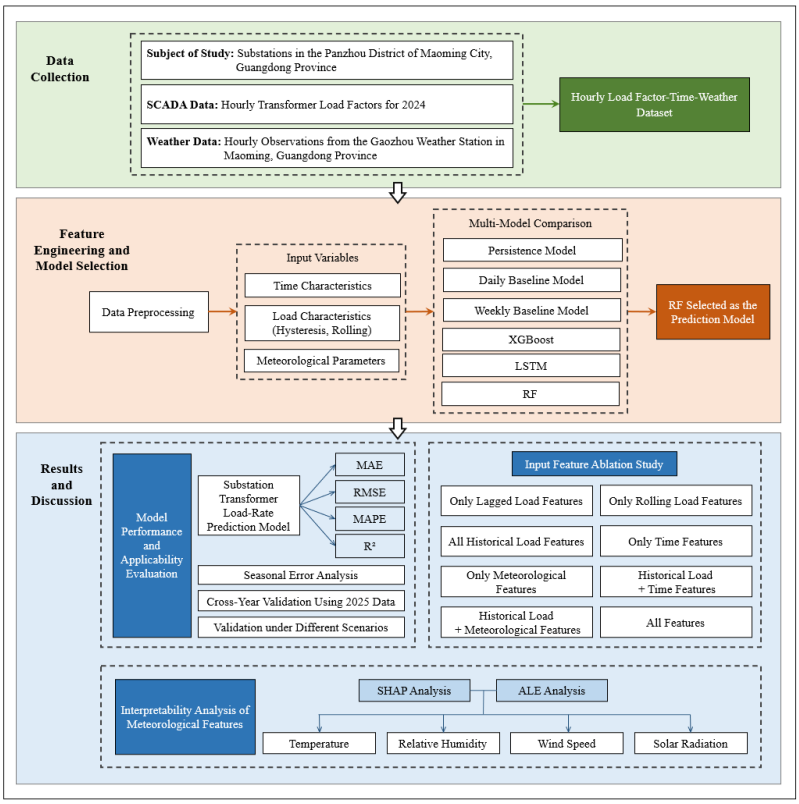


Figure 1. The Roadmap of the Study. Note: Arrows indicate the sequence of the research workflow and the connections between the main analytical steps.

2. Methodology

This section provides a brief description of the actual project and the data sources. It introduces the RF model, data preprocessing and feature engineering, benchmark model settings, and the hyperparameter tuning method based on rolling-origin validation. The SHAP and ALE methods used for model interpretation are also presented.

2.1. Data Sources

This study selected a 40 MVA main transformer at the Panzhou regional substation in Maoming City, Guangdong Province, China, as the research case. The substation is located at 110.80° E and 21.90° N. As a regional power supply hub, the substation primarily serves residential and industrial users, and its transformer load-rate exhibits pronounced fluctuations during operation. According to the operational records available for the study period, no major changes were known to have occurred in the main composition of electricity consumers, applicable electricity tariff mechanisms, or operating modes of the main transformer between 2024 and 2025. These conditions provided a relatively stable operational basis for cross-year model validation.

For data collection, the the Supervisory Control and Data Acquisition (SCADA) system of the substation automatically recorded the hourly load-rate of transformer data for the year 2024. In this study, the apparent power provided by the SCADA system is used as the indicator to represent the load. The load-rate of the substation transformers is defined as the unit loading level of the transformer relative to its rated apparent power capacity. Specifically, the load-rate is calculated as the ratio of the actual apparent power of the transformer to its rated capacity, and is expressed as a percentage. The mathematical expression is as follows:

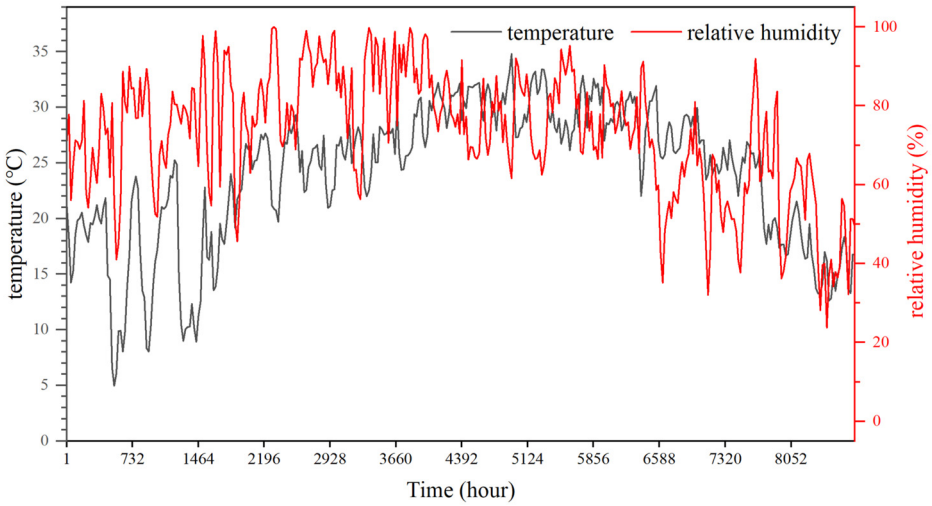
$$Load - rate(t) = \frac{S_{\text{actual}}(t)}{S_{\text{rated}}} \times 100\% \quad (1)$$

In the formula, $S_{\text{actual}}(t)$ denotes the apparent power measured by the substation's SCADA system at time t, while S_{rated} represents the rated capacity of the transformers. In this study, the main transformer under analysis has a rated capacity of 40 MVA.

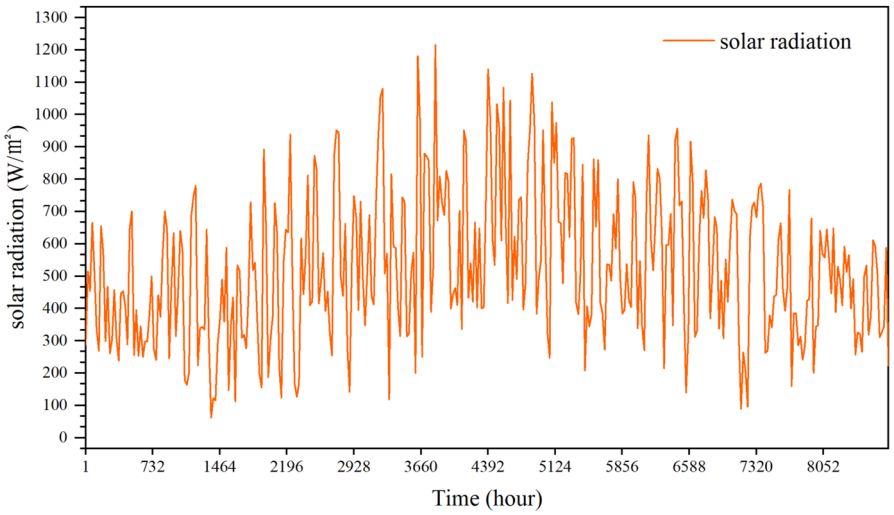
The meteorological parameters used to characterize the external weather conditions in the substation area were obtained from the China Meteorological Data Service Centre (CMDC; https://data.cma.cn (accessed on 1 March 2026)). The selected station was the Maoming Gaozhou Meteorological Station, located at 110.86° E and 21.94° N, approximately 5 km from the study substation in straight-line distance. To ensure temporal consistency between the meteorological and substation operating data, all meteorological records were converted to Beijing Time (UTC + 8) and matched with the hourly substation load-rate data recorded by the SCADA system according to their timestamps. The meteorological parameters included air temperature, relative humidity, wind speed, and solar radiation. Their units were standardized as °C, %, m/s, and W/m^2, respectively. Missing and anomalous meteorological records were first examined through temporal continuity and physical-range checks to identify missing timestamps, duplicate records, and values that were clearly inconsistent with physical constraints. A small number of missing values were imputed using linear interpolation or observations from adjacent hours. The hourly variations in air temperature, relative humidity, solar radiation, and wind speed during 2024 are presented in Figure 2.

The meteorological parameters used in this study were hourly observations recorded by the meteorological station. Model training, testing, and interpretability analyses were all conducted using these observed meteorological data, without incorporating weather forecast data. Therefore, this study addresses transformer load-rate forecasting conditional on observed meteorological conditions at the target time. The reported model performance

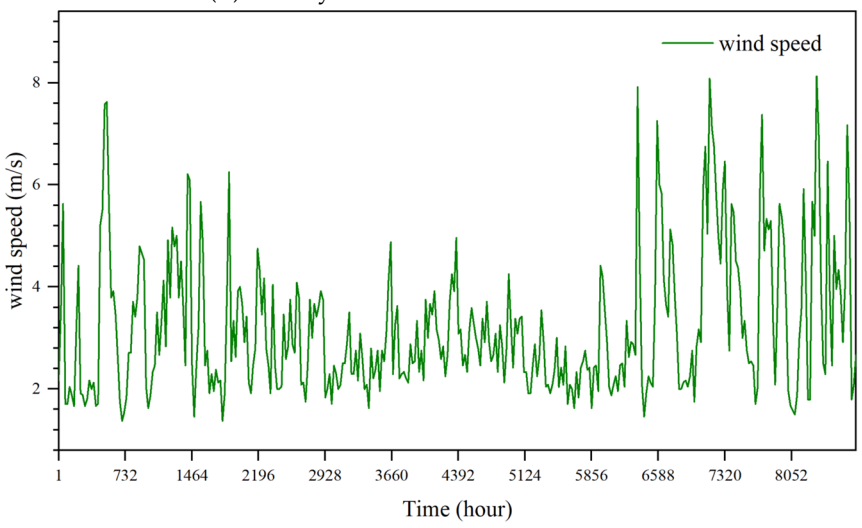
reflects forecasting accuracy under observed meteorological inputs, without accounting for potential weather forecast errors and their propagation in practical forecasting applications.



(**a**) Hourly Temperature and Humidity for 2024



(**b**) Hourly Solar Radiation for 2024



(**c**) Hourly Wind Speed for 2024

Figure 2. Actual meteorological data at the substation location.

2.2. Analytical Methods

2.2.1. RF Regression Model

RF is an ensemble learning algorithm that accomplishes classification and regression tasks by constructing multiple decision trees. The algorithm first employs the bootstrap method to repeatedly sample the original training set (of size N), generating multiple training subsets. Each subset contains n samples, allowing for sample repetition. Each decision tree is trained independently on a distinct subset, and only a randomly selected subset of features is used for node splitting. Finally, the RF model aggregates the outputs from the multiple decision trees through majority voting or averaging to produce the final prediction. The algorithmic structure is illustrated in Figure 3.

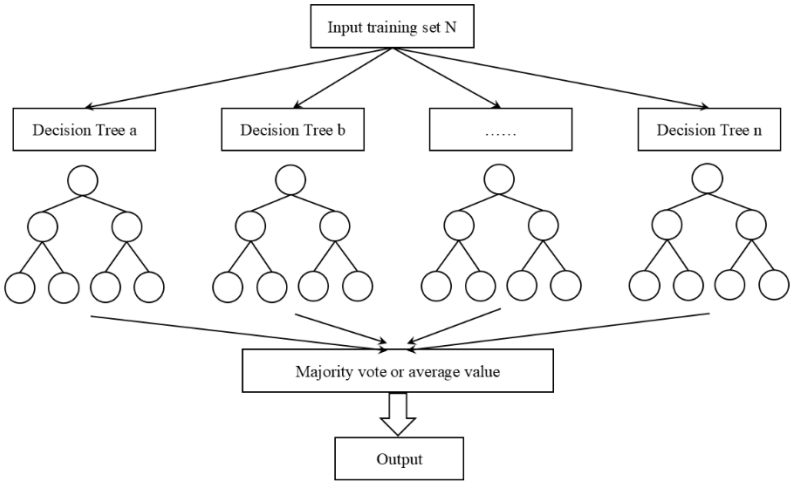


Figure 3. Structure Diagram of the RF Model. *Note: Circles represent decision-tree nodes, connecting lines indicate tree branches, and arrows illustrate the flow of data and the aggregation of tree predictions.*

Due to the simultaneous application of random sampling for both instances and features during training, RF model exhibits significantly enhanced generalization capabilities compared to single decision trees, effectively mitigating the risk of overfitting . Additionally, this algorithm demonstrates strong robustness against outliers and noisy data, making it particularly suitable for modeling scenarios involving complex inter-variable relationships or pronounced nonlinear characteristics. In model evaluation and error analysis, commonly used metrics include MAE, RMSE, and MAPE [26,27]. Furthermore, the coefficient of determination (R^2) is employed to measure the goodness of fit between predicted and actual values, where an R^2 value closer to 1 indicates better model fit [27]. The calculation formulas for each metric are as follows.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \in [0, +\infty) \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \in [0, +\infty) \tag{3}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \in [0, +\infty) \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \in (-\infty, 1] \tag{5}$$

The load-rate in this study is expressed as a percentage. The MAE, RMSE, Mean Bias Error (MBE), and Maximum Error (MaxE) represent absolute differences between the predicted and actual load-rates and are uniformly reported in percentage points. The MAPE

represents relative error and is expressed as a percentage, whereas the R^2 is dimensionless. When proportional values within the range of 0–1 are used during model computation, the MAE, RMSE, MBE, and MaxE are multiplied by 100 when reporting the results to convert them into percentage points.

2.2.2. Data Preprocessing and Feature Engineering

To clearly define the temporal boundary of the forecasting task, the substation load-rate at time (t), denoted as Load-rate(t), was specified as the prediction target. To predict the load-rate at the target time, the model used temporal, holiday, and meteorological features at time (t), together with historical load-rate features at (t-1), (t-24), and (t-168). All load-rate lag features and rolling statistics were calculated exclusively from historical load-rate data preceding time (t), without incorporating the actual load-rate at the target time, thereby preventing data leakage.

Before feature engineering, the dataset was preprocessed by standardizing the time formats of the load-rate and meteorological data, matching the two datasets according to hourly timestamps, and checking for missing records, duplicate observations, and anomalous values. A small number of missing values were imputed using linear interpolation or data from adjacent hours to ensure data completeness and temporal continuity.

To examine the statistical stationarity of the hourly load-rate series, the Augmented Dickey–Fuller (ADF) test [28] and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [29] were applied to the original load-rate series and its first-order differenced series. The null hypothesis of the ADF test is that the series contains a unit root, whereas the null hypothesis of the KPSS test is that the series is stationary. Both tests were conducted at a significance level of 5%. The combined results of these two tests were used to assess the stationarity characteristics of the load-rate series and provide a basis for subsequent data analysis in the forecasting model.

In load-rate forecasting, the original meteorological variables (temperature, relative humidity, wind speed, and solar radiation) can characterize changes in external environmental conditions but are insufficient to fully describe the periodic patterns and behavioral factors associated with temporal load variations. This is because regional electricity demand is associated not only with meteorological conditions but also with daily activity schedules, industrial operating cycles, differences between working days and public holidays, and load levels observed during the same period on the previous day or the previous week. Such periodic and behavioral information is not captured by the original meteorological data. Therefore, relying solely on meteorological variables may limit the model's ability to characterize the main factors associated with load variations and their temporal patterns.

Based on the above considerations, this study constructed multiple supplementary features in addition to the original meteorological variables, with the aim of enhancing the model's ability to represent temporal structures, behavioral patterns, and nonlinear relationships. The feature engineering procedures were as follows:

(1) Temporal Periodic Features

Based on the hourly data, actual timestamps were generated and the variables of hour, day-of-week, and month were extracted to characterize the periodic variations in residential energy use, commercial operations, and industrial production. Considering the 24 h periodicity of the hour variable, sine and cosine functions were applied for cyclical encoding to avoid placing hour 23 and hour 0 far apart in the numerical space despite being adjacent in actual time, thereby improving the model's ability to capture intra-day cyclical patterns.

(2) Holiday and Working-Day Features

Regional electricity consumption exhibited significant differences among working days, weekends, and official public holidays. Based on the 2024 national holiday calendar of China, indicator variables for holidays, weekends, and working days were constructed to reflect the impacts of institutionally driven behavioral changes on the electricity load. These features helped prevent the model from misinterpreting load jumps caused by institutional factors as meteorological effects, thereby reducing systematic bias.

(3) Lagged Load-Rate Features

The actual variation in the load-rate typically exhibited continuity and recurrence. For example, the load level at a given time tended to follow similar patterns to that of the previous hour, the same hour of the previous day, and the same hour of the previous week. Accordingly, this study constructed lagged features at t-1, t-24, and t-168 h, corresponding to short-term persistence, daily periodicity, and weekly periodicity, respectively. All lagged variables were generated by shifting the time series backward, ensuring that the model only used historical information and preventing leakage of future data.

(4) Rolling Statistical Features

To characterize the smoothed trends and fluctuations of the load over short- and medium-term periods, rolling means, rolling maxima, and rolling standard deviations were constructed using 24 h and 168 h windows. To prevent data leakage, the load-rate series was first shifted by 1 h before the rolling statistics were calculated from the shifted historical series. For a target time (t), the 24 h rolling window included only the historical load-rates from (t-24) to (t-1), whereas the 168 h rolling window included only those from (t-168) to (t-1). The latest endpoint of each rolling window was therefore (t-1), and the actual load-rate at the target time (t) was excluded. These statistical features provide information on local load trends without introducing future data and enhance the model's ability to represent different operating conditions.

(5) Nonlinear Expansion of Meteorological Variables

The relationships between meteorological variables and the load-rate typically exhibited strong nonlinear characteristics. For example, cooling demand increases sharply at high temperature levels, leading to accelerated growth in the load. To enhance the model's representational capacity under extreme meteorological conditions, squared terms of temperature, humidity, wind speed, and solar radiation were constructed to capture their nonlinear effects in high-value ranges.

After feature construction, Pearson correlation coefficients and the Variance Inflation Factor (VIF) were used for descriptive diagnostics of the correlation structure among input variables, with Spearman correlation coefficients employed as a supplementary measure [30]. Considering that the maximum historical window was 168 h, the first 168 h of data were excluded, yielding 8616 valid samples for 2024. Correlation diagnostics were conducted separately for three feature groups: the original meteorological variables, meteorological and historical load-rate features, and continuous features further incorporating the squared terms of meteorological variables. VIF values were calculated using auxiliary linear regression models with an intercept, excluding the target load-rate and temporal and calendar encoding variables. This analysis was primarily intended to identify linear correlations and potential redundancy among features in different groups, rather than to guide model parameter selection or establish causal relationships between variables.

Finally, the dataset was divided into training and testing sets in chronological order, ensuring that model training used only data preceding the corresponding test period to prevent the introduction of future information during time-series data splitting.

In addition, to evaluate the sensitivity of model forecasting performance to meteorological input errors, a meteorological input perturbation experiment was conducted using

the independent validation set from 2025. Taking the standard deviation of each meteorological variable in the 2024 training set as the reference scale, 5%, 10%, and 20% of the corresponding standard deviation were used to define low-, medium-, and high-intensity perturbation levels, respectively. Simulated meteorological inputs were generated by introducing zero-mean random perturbations. A total of eight scenarios were considered, including an unperturbed baseline, four single-variable scenarios with medium-intensity perturbations, and three joint perturbation scenarios with different intensity levels. While keeping the trained RF model, historical load features, and temporal features unchanged, the corresponding perturbations were applied to temperature, relative humidity, wind speed, and solar radiation, with the meteorologically derived features updated accordingly. The sensitivity of model forecasting performance to meteorological input errors was evaluated by comparing MAE, RMSE, MAPE, and R^2 across the different scenarios.

2.2.3. Benchmark Model Setup and Hyperparameter Tuning Using Time-Series Validation

To evaluate the applicability of different forecasting models to hourly transformer load-rate forecasting at regional substations, this study selected a persistence model, a daily periodic baseline, a weekly periodic baseline, LSTM, XGBoost, LightGBM, and CatBoost as comparison models, establishing a comparative framework encompassing time-series baselines, deep learning, and tree-based ensemble methods. Specifically, the persistence model used the load-rate from the previous hour as the prediction for the current time, while the daily and weekly periodic baselines used the load-rate at the corresponding hour 24 h and 168 h earlier, respectively, to assess the predictive information provided by temporal continuity and periodicity in the load-rate series. LSTM was employed to evaluate the predictive performance of deep sequence learning methods, while XGBoost, LightGBM, and CatBoost were selected as representative gradient-boosted tree algorithms for comparison with the RF model. All models were evaluated using the same dataset, feature-processing procedures, chronological data-splitting scheme, and evaluation metrics to examine their forecasting performance and computational costs under the conditions considered in this study.

Data preprocessing, feature engineering, and model development were performed using Python 3.12.7, with NumPy 1.26.4 and pandas 2.2.2. The RF, XGBoost, LightGBM, CatBoost, and LSTM models were implemented using scikit-learn 1.5.1, XGBoost 2.0.3, LightGBM 4.6.0, CatBoost 1.2.10, and TensorFlow 2.19.0 with Keras 3.10.0, respectively. SHAP analysis was conducted using SHAP 0.50.0, while ALE analysis was implemented using custom Python functions. Statistical analyses and visualization were performed using SciPy 1.13.1, statsmodels 0.14.2, and Matplotlib 3.9.2.

To reduce the potential risk of information leakage during time-series data splitting, rolling-origin validation was employed for model hyperparameter tuning and performance evaluation. Consecutive training windows and subsequent test windows were arranged in chronological order. In each iteration, the models were trained exclusively on data preceding the corresponding test window and evaluated on the subsequent time window. For RF, XGBoost, LSTM, LightGBM, and CatBoost, hyperparameters were selected based on the mean root mean square error (RMSE) across the corresponding rolling validation windows, and the parameter configurations for each model were determined accordingly. The hyperparameter search ranges and final parameter values for each model are presented in Table 1.

To balance the training sample size, test-window length, and completeness of seasonal coverage, nine rolling test windows were used to evaluate the forecasting performance of all models under the same validation scheme. The means and standard deviations of mean absolute error (MAE), RMSE, mean absolute percentage error (MAPE), and coefficient of

determination (R^2) across the test windows were calculated to assess overall predictive performance and variability across different windows. Table 2 presents the forecasting performance of the different models across the nine rolling test windows in 2024.

Table 1. Parameter Settings and Hyperparameter Search Ranges for the Benchmark Models.

Model	Parameters	Search Range/Model Setting	Final Value
Persistence	lag	1	$\hat{y}_t = y_{t-1}$
Daily baseline	lag	24	$\hat{y}_t = y_{t-24}$
Weekly baseline	lag	168	$\hat{y}_t = y_{t-168}$
XGBoost	n_estimators	{100, 200, 300, 500}	300
	max_depth	{3, 5, 7, 9}	7
	learning_rate	{0.01, 0.03, 0.04, 0.05, 0.1}	0.04
	subsample	{0.7, 0.8, 0.85, 0.9, 1.0}	0.85
	colsample_bytree	{0.7, 0.8, 0.85, 0.9, 1.0}	0.85
LSTM	input sequence length	[1, 24, 48, 72, 168]	1
	number of hidden units	[32, 64, 128, 256]	128
	layers	[1, 2]	1
	Dropout	[0.1, 0.15, 0.2, 0.3]	0.1
	learning_rate	[0.0001, 0.001, 0.01, 0.1]	0.001
	Batch size	[32, 64]	64
RF	n_estimators	100, 200, 300, 500	200
	max_depth	None, 10, 20, 30	None
	random_state	Fixed Value	42
LightGBM	n_estimators	{100, 200, 300}	300
	max_depth	{10, 12, 15, −1}	15
	learning_rate	{0.03, 0.05, 0.10}	0.03
	num_leaves	{31}	31
	random_state	Fixed-Value	42
CatBoost	iterations	{100, 200}	200
	depth	{6, 8}	8
	learning_rate	{0.05, 0.10}	0.10
	random_seed	Fixed-Value	42

Note: Friedman tests were conducted across five machine learning models (RF, LSTM, XGBoost, LightGBM, and CatBoost) using nine paired rolling test windows. Following a significant overall test, two-sided Wilcoxon signed-rank tests were performed using RF as the reference model, with Holm correction applied separately for each evaluation metric across four pairwise comparisons. A Holm-adjusted p-value below 0.05 indicates a statistically significant pairwise difference.

As shown in Table 2, the forecasting performance of the models varied under rolling-origin validation. The persistence model and daily periodic baseline captured the short-term continuity and daily periodicity of the load-rate series. In contrast, the weekly periodic baseline achieved mean MAE, RMSE, and MAPE values of 4.56 percentage points, 5.46 percentage points, and 13.29%, respectively, with a mean R^2 of only 0.2280. These metrics also exhibited substantial variability across the test windows, indicating that relying solely on the load-rate at the same hour in the previous week was insufficient to consistently characterize hourly load-rate variations. Among the machine learning models, LSTM achieved mean RMSE and R^2 values of 3.00 percentage points and 0.8685, respectively, outperforming XGBoost in these two metrics. In contrast, XGBoost achieved a mean MAE of 2.08 percentage points and a mean MAPE of 5.56%, outperforming LSTM in the corresponding metrics. Among the tree-based ensemble models, LightGBM achieved relatively low mean prediction errors, with MAE, RMSE, and MAPE values of 1.518 percentage points, 2.326 percentage points, and 4.178%, respectively, together with a mean R^2 of 0.9137, yielding the most favorable mean numerical results across all four metrics among the comparison models. The RF model achieved mean MAE, RMSE, and MAPE values of 1.60 percentage

points, 2.41 percentage points, and 4.45%, respectively, with a mean R^2 of 0.9087, indicating forecasting performance relatively close to that of LightGBM. CatBoost achieved mean MAE and RMSE values of 1.766 and 2.626 percentage points, respectively, with a MAPE of 4.901%, exhibiting overall performance between RF and XGBoost. Further comparison of performance variability across the test windows showed that LightGBM exhibited certain numerical advantages in mean forecasting accuracy, whereas RF had relatively smaller standard deviations in RMSE, MAPE, and R^2 across the test windows. These findings indicate that the tree-based ensemble models exhibited differences in forecasting accuracy and performance variability across the test windows. However, the statistical significance of these performance differences requires further examination.

Table 2. Performance Variability of Different Models under Rolling-Origin Validation in 2024 (Mean ± Standard Deviation).

Model	MAE(Percentage Points)	RMSE(Percentage Points)	MAPE(%)	R^2
Persistence	2.40 ± 0.23	3.04 ± 0.38	7.88 ± 1.84	0.8426 ± 0.0342
Daily baseline	2.33 ± 1.33	3.29 ± 2.00	6.69 ± 2.80	0.8011 ± 0.1848
Weekly baseline	4.56 ± 3.61	5.46 ± 4.03	13.29 ± 10.52	0.2280 ± 1.3517
LSTM	2.21 ± 0.65	3.00 ± 0.85	6.57 ± 1.37	0.8685 ± 0.0484
XGBoost	2.08 ± 1.30	3.06 ± 2.09	5.56 ± 2.23	0.8377 ± 0.2010
RF	1.60 ± 0.65	2.41 ± 1.17	4.45 ± 0.85	0.9087 ± 0.0733
LightGBM	1.518 ± 0.691	2.326 ± 1.274	4.178 ± 0.977	0.9137 ± 0.0810
CatBoost	1.766 ± 0.900	2.626 ± 1.609	4.901 ± 1.521	0.8869 ± 0.1250

To further assess the performance differences among the machine learning models under rolling-origin validation, the nine rolling test windows were treated as paired samples. Friedman tests were conducted separately for the MAE, RMSE, MAPE, and R^2 values of five machine learning models: RF, LSTM, XGBoost, LightGBM, and CatBoost. When the overall Friedman test reached the significance level of 0.05, two-sided Wilcoxon signed-rank tests were further conducted for pairwise comparisons using RF as the reference model, followed by Holm correction for multiple testing. The statistical test results are presented in Table 3.

As shown in Table 3, the Friedman statistics for MAE, RMSE, MAPE, and R^2 across the five machine learning models were 28.1341, 18.4000, 26.6592, and 19.5556, respectively, with all corresponding p-values below 0.05, indicating statistically significant overall differences in forecasting performance among the models across the rolling test windows. Further Wilcoxon signed-rank tests showed that, for MAE and MAPE, the differences between RF and LSTM, XGBoost, and LightGBM were statistically significant (Holm-adjusted $p < 0.05$), whereas the differences between RF and CatBoost were not significant. Combined with the mean prediction errors presented in Table 2, these results indicate that RF outperformed LSTM and XGBoost in terms of MAE and MAPE, while LightGBM achieved better performance than RF in these two metrics. For RMSE, only the difference between RF and LightGBM reached statistical significance (Holm-adjusted $p = 0.0469$), while the remaining pairwise comparisons were not statistically significant. For R^2, although the overall Friedman test indicated significant differences among the five models, none of the pairwise Wilcoxon comparisons reached the 0.05 significance level after Holm correction.

These results indicate that LightGBM exhibited statistically significant performance advantages in certain prediction error metrics under the rolling-origin validation scheme considered in this study, while RF maintained relatively strong predictive performance in terms of MAE and MAPE compared with LSTM and XGBoost. The performance differences among the models were metric-dependent, suggesting that model evaluation should

comprehensively consider prediction errors, performance variability across test windows, and statistical significance.

Table 3. Statistical significance tests of the rolling-origin validation results for the machine-learning models.

Evaluation Metric	Friedman Statistic	Friedman *p*-Value	Pairwise Comparison	Wilcoxon Statistic	Raw *p*-Value	Holm-Adjusted *p*-Value
MAE	28.1341	0.000012	RF–LSTM	0	0.0039	0.0156
			RF–XGBoost	0	0.0078	0.0156
			RF–LightGBM	0	0.0039	0.0156
			RF–CatBoost	16	0.4961	0.4961
RMSE	18.4000	0.001031	RF–LSTM	6	0.0547	0.1094
			RF–XGBoost	3.5	0.0234	0.0703
			RF–LightGBM	2	0.0117	0.0469
			RF–CatBoost	20	0.8203	0.8203
MAPE	26.6592	0.000023	RF–LSTM	0	0.0039	0.0156
			RF–XGBoost	1	0.0156	0.0313
			RF–LightGBM	0	0.0039	0.0156
			RF–CatBoost	16	0.4961	0.4961
R^2	19.5556	0.000611	RF–LSTM	6	0.0547	0.1094
			RF–XGBoost	4	0.0273	0.0820
			RF–LightGBM	3	0.0195	0.0781
			RF–CatBoost	20	0.8203	0.8203

After completing rolling-origin validation and hyperparameter tuning, the models were further evaluated on the independent test set for 2024 using the selected parameter configurations. The comparative forecasting results are presented in Table 4.

Table 4. Forecasting Performance of Different Models on the Independent Test Set for 2024.

Model	MAE (Percentage Points)	RMSE (Percentage Points)	MAPE (%)	R^2
Persistence	2.41	3.07	8.04	0.9419
Daily baseline	2.27	3.79	6.58	0.9117
Weekly baseline	4.48	6.73	13.08	0.7221
LSTM	1.43	4.87	6.09	0.8934
XGBoost	1.48	2.07	4.21	0.9541
RF	0.97	1.46	3.93	0.9654
LightGBM	0.915	1.337	3.674	0.9710
CatBoost	1.848	2.218	8.943	0.9203

As shown in Table 4, the forecasting performance of the models varied on the independent test set for 2024. The weekly periodic baseline exhibited relatively large prediction errors, with a MAPE of 13.08% and an R^2 of 0.7221. The persistence model and daily periodic baseline achieved R^2 values of 0.9419 and 0.9117, respectively, indicating that recent historical load-rate and daily periodic information provided useful references for forecasting. Among the machine learning models, LightGBM achieved an MAE of 0.915 percentage points, an RMSE of 1.337 percentage points, a MAPE of 3.674%, and an R^2 of 0.9710, yielding relatively favorable numerical results among the comparison models. The RF model achieved an MAE of 0.97 percentage points, an RMSE of 1.46 percentage points, a MAPE of 3.93%, and an R^2 of 0.9654, indicating forecasting performance relatively close

to that of LightGBM. XGBoost achieved a MAPE of 4.21% and an R^2 of 0.9541, whereas CatBoost achieved corresponding values of 8.943% and 0.9203, indicating differences in forecasting performance among the gradient-boosted tree algorithms on this test set. Overall, LightGBM and RF exhibited relatively low prediction errors on the independent test set examined in this study. However, these results reflect numerical comparisons under the current dataset and model configurations and should be evaluated in conjunction with the rolling-origin validation results.

Considering the results of rolling-origin validation and independent testing, both LightGBM and RF demonstrated relatively strong predictive performance under the dataset and feature system considered in this study. LightGBM achieved relatively favorable numerical results in terms of prediction errors and R^2, whereas RF exhibited smaller standard deviations in RMSE, MAPE, and R^2 across the rolling test windows. Given the relatively strong predictive capability of RF under the data conditions of this study and its suitability for conducting feature ablation experiments and SHAP and ALE analyses using a fixed trained model, RF was selected as the primary model for subsequent analyses of meteorological feature contributions and model response relationships.

To further evaluate the computational costs of different machine learning models, the training time and peak memory usage of RF, XGBoost, LSTM, LightGBM, and CatBoost were compared under the same computing environment, as presented in Table 5. XGBoost required the shortest training time of 3.0205 s, followed by RF and LightGBM at 4.3504 s and 4.4256 s, respectively, with relatively similar training times. CatBoost required 5.7615 s, while LSTM had a relatively longer training time of 9.9077 s. In terms of memory usage, RF had the lowest peak memory usage of 4.5789 MB, whereas XGBoost, LightGBM, and CatBoost required 9.3730, 10.8262, and 12.1145 MB, respectively. LSTM exhibited the highest peak memory usage of 30.3554 MB. Combined with the forecasting performance results, LightGBM demonstrated relatively high predictive accuracy at the dataset scale considered in this study, while RF exhibited comparatively low memory usage. These findings indicate that the models differed in forecasting performance and computational resource requirements, providing a reference for subsequent model selection.

Table 5. Training Costs of Different Models.

Model	Training Time (s)	Peak Memory Usage (MB)
RF	4.3504	4.5789
XGBoost	3.0205	9.3730
LSTM	9.9077	30.3554
LightGBM	4.4256	10.8262
CatBoost	5.7615	12.1145

2.2.4. SHAP Algorithm

To quantify the contributions of different input features to the predictions of the RF model, Shapley Additive Explanations (SHAP) was employed for model interpretability analysis [31]. Derived from Shapley values in cooperative game theory, SHAP calculates the marginal contribution of each feature to the model output across different feature combinations. It assigns feature attribution values to individual samples, which can then be aggregated to assess global feature importance and contribution direction. Compared with permutation feature importance, which primarily measures changes in overall predictive performance after feature permutation, SHAP provides both local and global explanations. Unlike Local Interpretable Model-agnostic Explanations (LIME), which uses local surrogate models to approximate model predictions, SHAP provides a unified feature attribution framework that is more closely aligned with the objective of analyzing the contributions of

meteorological features to model predictions in this study. The calculation formula is as follows:

$$\varphi_i = \sum_{S \subseteq N(i)} \frac{|S|!(n-|S|-1)!}{n!}[f(S \cup \{i\}) - f(S)] \quad (6)$$

In the formula, φ_i is the SHAP value for feature i; N denotes the set of all features; S is a subset of features excluding feature i; $|S|$ is the number of elements in set S; n is the total number of elements in set N; and $f(S \cup \{i\}) - f(S)$ represents the incremental change in the model output when feature i is added to the subset S. In this computation, each sample corresponds to a set of SHAP values, reflecting the individual contribution of each input variable to the model output for that sample. A larger absolute SHAP value indicates a greater magnitude of the feature's attribution contribution to the model prediction for the corresponding sample.

To analyze the relative importance and contribution direction of meteorological features in model predictions, mean absolute SHAP value bar charts and SHAP summary plots were used for visualization. The mean absolute SHAP value measures the average magnitude of feature attribution across the evaluation samples, while the SHAP summary plot illustrates the distribution of feature attribution values across individual samples and their relationships with feature values.

SHAP values represent the contributions of individual features to model predictions under a given model and data distribution, rather than causal relationships between features and load-rate. When input features are correlated, attribution contributions may be distributed among correlated features. Therefore, the SHAP results in this study were primarily used to compare the relative contributions of meteorological variables to model predictions and their contribution directions.

To evaluate the statistical stability of meteorological feature attributions, 1000 Bootstrap resampling iterations were performed on the independent test set using the fixed RF model. The empirical distributions of the mean absolute SHAP values for each meteorological variable were calculated, and 95% empirical intervals were constructed using the 2.5th and 97.5th percentiles. By comparing the attribution distributions and interval ranges of different meteorological variables, the sensitivity of SHAP feature importance results to variations in the evaluation samples was assessed. This analysis reflects the stability of feature attributions under the fixed model and current evaluation samples, without accounting for variations in model interpretations arising from model retraining.

2.2.5. ALE Analysis

SHAP can quantify the attribution contributions of input features to model predictions. However, feature importance rankings alone cannot fully characterize how model predictions vary with changes in meteorological variables. To further analyze the nonlinear responses of the model to meteorological variables, Accumulated Local Effects (ALE) was employed as a supplementary interpretability method.

The commonly used Partial Dependence Plot (PDP) describes the relationship between a target feature and model predictions by averaging over other features. However, when input features are correlated, PDP may involve feature combinations that are rarely observed in the actual data, potentially reducing the reliability of the interpretation. In contrast, ALE calculates and accumulates local differences in model predictions between adjacent feature intervals within the observed data distribution, thereby reducing, to some extent, the extrapolation problems associated with unrealistic feature combinations under feature correlation. Individual Conditional Expectation (ICE) primarily focuses on displaying prediction responses for individual samples. Since this study mainly examines the overall model response trends associated with meteorological variables within the observed sample range, ALE was selected for the analysis.

ALE calculates differences in model predictions when the target feature varies between the boundaries of adjacent intervals and accumulates the average local effects across these intervals to characterize how model predictions vary with feature values. The calculation formula is as follows:

$$\tilde{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_{k,j}(x)} \sum_{i:x_j^{(i)} \in N_{k,j}(x)} \left[f(z_{k,j}, x_j^{(i)}) - f(z_{k-1}, x_j^{(i)}) \right] \quad (7)$$

Here, j denotes the target feature index, k represents the bin index, and $k_j(x)$ denotes the index of the bin containing the target feature value (x). $N_{k,j}(x)$ represents the k-th interval of feature j, with $n_{k,j}(x)$ observations and lower and upper boundaries $z_{k-1,j}(x)$ and $z_{k,j}(x)$, respectively. The function $\tilde{f}$ denotes the trained RF prediction model. For each interval, the local prediction differences are averaged and subsequently accumulated to obtain the uncentered ALE values. The ALE curve is then centered by subtracting the sample mean of the accumulated effects. In this study, the corresponding squared meteorological feature was updated simultaneously when its original feature value was varied. Positive ALE values indicate relatively higher model-predicted load-rates compared with the centered reference level, rather than a causal effect on actual load-rate. The ALE values were converted to percentage points for interpretation.Quantile-based equal-frequency binning was applied to the meteorological variables, with the baseline number of bins set to 20 to ensure a relatively balanced sample distribution across intervals. Considering that the number of bins may affect the local shapes of ALE curves, a sensitivity analysis was further conducted using 10, 20, and 30 bins. Taking the results obtained with 20 bins as the reference, Pearson correlation coefficients between ALE response curves were calculated to compare the overall trends under different binning configurations and assess the sensitivity of the interpretation results to the number of bins.

ALE reflects the average local model response under a given model and observed data distribution. Although ALE can reduce, to some extent, the extrapolation problems associated with feature correlation compared with PDP, it cannot completely eliminate potential interpretation bias arising from feature dependencies. Therefore, ALE was primarily used in this study to analyze the nonlinear response relationships between meteorological variables and model predictions, rather than to establish independent causal effects of meteorological factors on actual transformer load-rate.

3. Results and Discussion

3.1. The RF Model

3.1.1. The Prediction Results of the Model

Before model training, the ADF and KPSS stationarity tests, together with the Pettitt test for potential change points, were conducted on the hourly load-rate series for 2024. The results are presented in Tables 6 and 7.

Table 6. Stationarity tests for the hourly transformer load-rate series.

Series	ADF Statistic	*p*-Value	KPSS Statistic	*p*-Value
Original load-rate	−2.9850	0.0363	4.0687	<0.01
First difference	−17.2999	<0.001	0.0158	>0.10

Table 7. Pettitt test for the 2024 load-rate series.

Series	Observations	Pettitt K	Change-Point Position	*p*
2024 load-rate	8784	907,880	Observation 3466	<0.001

The ADF and KPSS tests did not provide consistent conclusions regarding the stationarity of the original load-rate series, whereas the first-order differenced series passed both tests, indicating clearer evidence of stationarity after differencing. In addition, the Pettitt test identified a statistically significant change point at the 3466th observation ($p < 0.001$), with mean load-rates of 26.49% and 37.75% before and after the change point, respectively. Considering that this change point occurred during the transition from spring to summer, the available evidence is insufficient to attribute it to changes in equipment or operating modes. Therefore, the original load-rate series was retained as the prediction target, and chronological data splitting and rolling-origin validation were employed to evaluate model performance.

Based on the established feature system and data preprocessing procedures, a forecasting dataset was constructed using hourly load-rate data from 2024, incorporating temporal, holiday, historical load-rate, rolling statistical, and meteorological features. The original time series contained 8784 hourly records. After excluding the first 168 h of data for which complete historical features could not be constructed, 8616 valid samples were obtained. The dataset was divided into training and testing sets in chronological order, with the period from 8 January to 20 October 2024 used for training and 21 October to 31 December 2024 used for testing (Figure 4). Model performance was then evaluated using the selected parameter configuration.

Figure 5 presents the comparison between the actual and predicted hourly load-rate of the substation obtained from the RF regression model driven by meteorological parameters, for the last 1728 hourly time points from 21 October to 31 December. This provided an intuitive visualization of the model's predictive performance. As shown in the figure, the RF model closely tracked the hourly variations in load-rate on the test set when observed meteorological parameters at the target time were used as inputs. The RF model achieved an R^2 of 0.9654, a MAPE of 3.93%, an MAE of 0.97 percentage points, and an RMSE of 1.46 percentage points, indicating relatively strong predictive performance under the conditional forecasting task defined in this study.

The above results indicate that the RF model achieved relatively strong hourly load-rate forecasting performance on the independent test set at the study site, providing a reference for analyzing transformer operating conditions. To further evaluate the uncertainty of model predictions, split conformal prediction was employed to construct 95% prediction intervals. On the test set, the prediction interval coverage probability (PICP) was 98.43%, with a mean prediction interval width (MPIW) of 9.98 percentage points. In addition, the average probabilities of the load-rate exceeding the thresholds of 50% and 60% were estimated from the empirical distribution of calibration residuals, yielding values of 1.343% and 0.232%, respectively. By combining point predictions of load-rate, prediction intervals, and threshold exceedance probabilities, additional information can be provided for assessing operating conditions in terms of predicted load levels, prediction uncertainty ranges, and potential high-load risks. Actual load allocation and operational dispatch decisions still require careful consideration of equipment capacity, operating constraints, and power supply reliability. The prediction results themselves do not directly correspond to specific load adjustment strategies or economic costs.

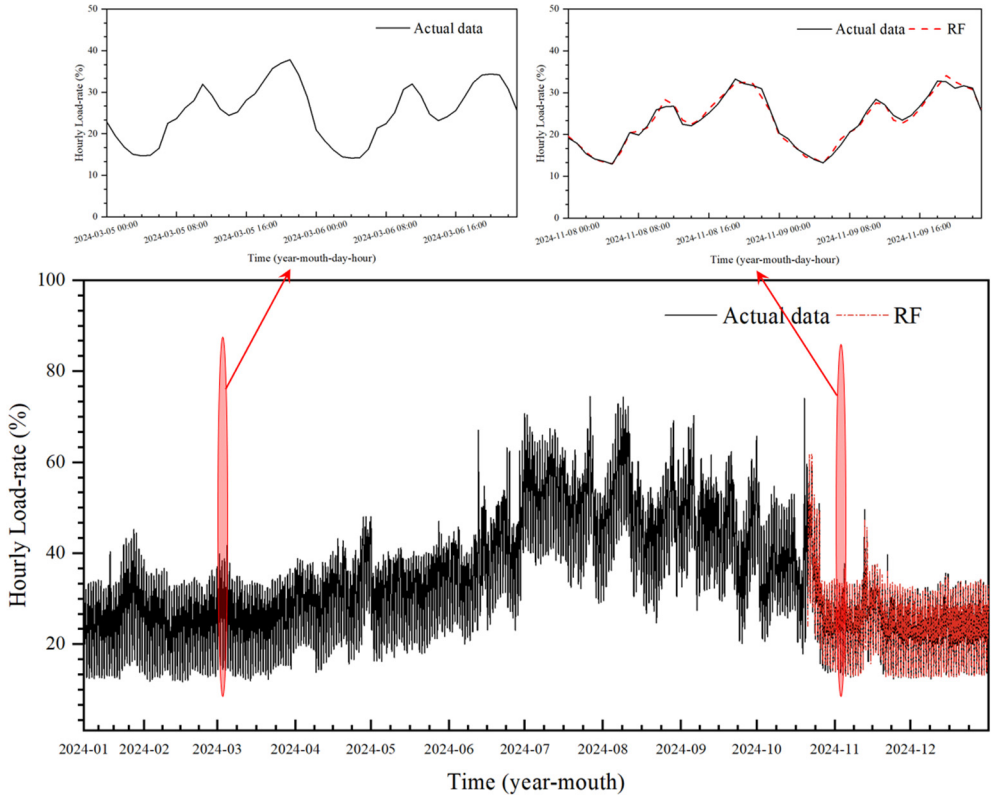


Figure 4. Actual vs. Forecast Hourly Load-rate for Substation.

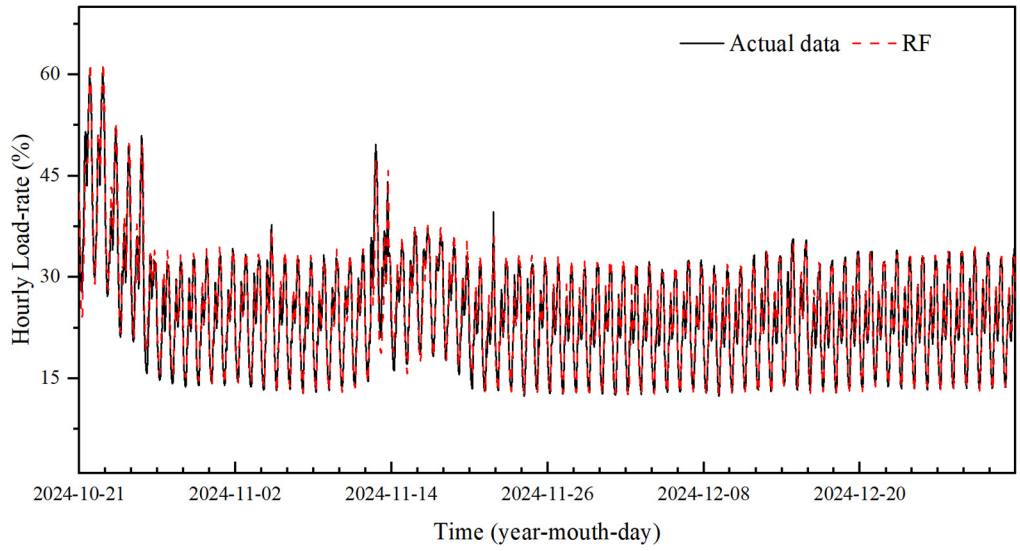


Figure 5. Hourly Forecast Results Comparison for Substation Load-rate.

3.1.2. Seasonal Error Analysis of the Model Based on Rolling-Window Testing

To provide a more comprehensive assessment of model performance across different seasons and examine whether high forecasting accuracy could be maintained under seasonal climatic variations, a seasonal error analysis was conducted. To avoid overly optimistic seasonal performance estimates caused by in-sample evaluation, the analysis was based on rolling-window time-series validation using the full-year 2024 dataset. Specifically, the model was trained and tested sequentially, with each iteration using only historical data preceding the current test window for model training and predicting the subsequent test window. The samples from all rolling test windows were then classified into spring, summer,

autumn, and winter according to their timestamps, and the MAE, RMSE, MAPE, and R^2 were calculated separately for each season.

Table 8 presents the seasonal errors of the model calculated from the rolling-window test samples. The model maintained a certain level of forecasting accuracy across all four seasons, although some differences were observed among them. The seasonal MAPE values ranged from 3.26% to 5.24%, while the R^2 values ranged from 0.8654 to 0.9741, indicating that the model generally adapted well to seasonal meteorological conditions and load-rate variation patterns throughout 2024. Forecasting performance was relatively better in winter and autumn, whereas lower accuracy was observed in summer.

Table 8. Seasonal Errors of the RF Model on Rolling Test Samples in 2024.

Season	MAE (Percentage Points)	RMSE (Percentage Points)	MAPE (%)	R^2
Spring	1.38	2.33	4.64	0.9107
Summer	2.50	3.97	5.24	0.8654
Autumn	1.43	2.14	4.24	0.9705
Winter	0.75	1.02	3.26	0.9741

In terms of absolute errors, winter showed the lowest errors, with an MAE of 0.75 percentage points, an RMSE of 1.02 percentage points, a MAPE of 3.26%, and an R^2 of 0.9741, indicating relatively stable forecasting performance during winter. In autumn, the MAE and RMSE were 1.43 percentage points and 2.14 percentage points, respectively, with a MAPE of 4.24% and an R^2 of 0.9705, also demonstrating strong predictive stability. By comparison, the errors were relatively larger in summer, with an MAE of 2.50 percentage points, an RMSE of 3.97 percentage points, a MAPE of 5.24%, and an R^2 of 0.8654. These results suggest that forecasting became more difficult under summer conditions characterized by high temperatures and greater load-rate variability, resulting in lower predictive accuracy than in the other seasons. In spring, the MAE and RMSE were 1.38 percentage points and 2.33 percentage points, respectively, with a MAPE of 4.64% and an R^2 of 0.9107. Overall, spring performance was intermediate between summer and autumn–winter, indicating that the model retained a certain degree of adaptability during the transitional season.

Based on the 2024 rolling-window time-series validation results, the model showed no evident seasonal failure and maintained relatively stable forecasting performance across all four seasons. Performance was strongest in winter and autumn, followed by spring, whereas summer exhibited comparatively higher errors. These findings indicate that the model has a certain degree of within-year temporal extrapolation capability and seasonal adaptability, although further improvement is still needed under high-temperature summer conditions and periods of pronounced load-rate variability.

3.1.3. Cross-Year Model Performance Under Different Conditions

To further evaluate the forecasting performance of the RF model on data from different years, the model was trained and its parameters were determined using data from 2024. Hourly load-rate data and corresponding meteorological observations from June to August 2025 were then used as an independent validation set. This period corresponds to summer in the study region, characterized by relatively high temperatures and pronounced load fluctuations, providing an opportunity to assess the model's predictive performance under cross-year warm-season conditions. This cross-year validation was limited to operational data from the same substation in different years, and the evaluation results primarily reflect

the model's temporal generalization capability under warm-season conditions at the same study site.

Model performance was evaluated using the complete independent validation dataset from June to August 2025. Figure 6 compares the actual and predicted values of the RF model on the 2025 validation set. The RF model closely tracked the daily periodic fluctuations and stage-specific variation trends in the hourly load-rate, with the predicted curve showing a generally high degree of agreement with the actual curve.

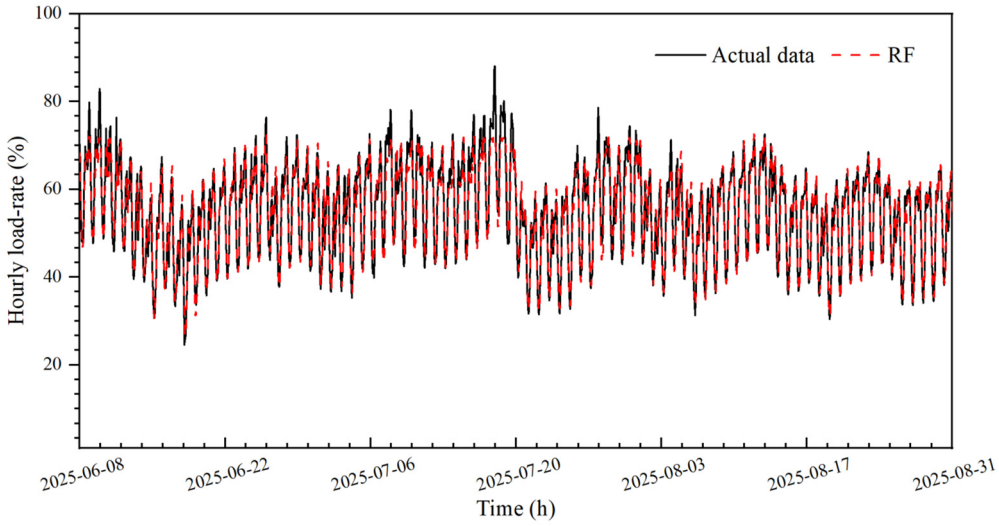


Figure 6. Comparison of Hourly Prediction Results on the Validation Set.

The forecasting accuracy, overall bias, and peak-load tracking capability of the RF model on the cross-year validation set were comprehensively evaluated using MAE, RMSE, MAPE, R^2, MBE, and MaxE. On the validation set from June to August 2025, the RF model achieved an MAE of 2.08 percentage points, an RMSE of 2.80 percentage points, a MAPE of 3.73%, an R^2 of 0.9291, an MBE of 0.37 percentage points, and a MaxE of 16.34 percentage points. These results indicate that the RF model maintained relatively strong predictive performance on the independent validation set from June to August 2025, with relatively low overall prediction errors. The MBE was close to zero, suggesting no evident systematic prediction bias. Meanwhile, the model was able to track the hourly load-rate fluctuations in the validation set to a certain extent, indicating that it retained a certain degree of predictive capability on cross-year warm-season data from the same substation. These findings reflect the model's temporal extrapolation performance under comparable seasonal conditions across different years.

To further evaluate changes in model forecasting performance under meteorological input errors, a sensitivity analysis of meteorological input perturbations was conducted using the independent validation set from 2025. The results are presented in Table 9.

Under the unperturbed baseline scenario, the model achieved MAE, RMSE, and MAPE values of 2.0803 percentage points, 2.8021 percentage points, and 3.7240%, respectively, with an R^2 of 0.929614. After applying medium-intensity perturbations to individual meteorological variables, the RMSE values ranged from 2.8321 to 2.8331 percentage points, while the R^2 values ranged from 0.927461 to 0.927515. Among these scenarios, the solar radiation perturbation scenario exhibited a relatively higher RMSE, although the differences in forecasting performance among the single-variable perturbation scenarios were relatively small.

Table 9. Prediction Performance of the RF Model under Different Meteorological Input Perturbation Scenarios.

Perturbation Scenario	MAE (pp)	RMSE (pp)	MAPE (%)	R^2
Unperturbed Baseline	2.08	2.80	3.72	0.9296
Temperature Only (Moderate)	2.12	2.83	3.82	0.9275
Relative Humidity Only (Moderate)	2.12	2.83	3.82	0.9275
Wind Speed Only (Moderate)	2.12	2.83	3.82	0.9275
Solar Radiation Only (Moderate)	2.12	2.83	3.82	0.9274
Combined Perturbation (Low)	2.12	2.83	3.82	0.9274
Combined Perturbation (Moderate)	2.12	2.83	3.82	0.9274
Combined Perturbation (High)	2.12	2.83	3.83	0.9273

Note: The unperturbed baseline uses observed meteorological inputs. Individual perturbation scenarios apply moderate perturbations to one meteorological variable at a time, while combined scenarios perturb all four meteorological variables simultaneously. Low, moderate, and high perturbation levels are defined as 5%, 10%, and 20% of the corresponding meteorological variable's standard deviation in the 2024 training dataset, respectively. The perturbations are hypothetical input-error scenarios rather than errors derived from operational weather forecasts.

Under joint perturbation conditions, as the perturbation intensity increased from low to high, MAE increased from 2.1205 to 2.1229 percentage points, RMSE increased from 2.8332 to 2.8360 percentage points, and MAPE increased from 3.8243% to 3.8288%, while R^2 decreased from 0.927458 to 0.927314, indicating a slight overall decline in forecasting performance. Under the high-intensity joint perturbation scenario, RMSE increased by approximately 1.21% compared with the unperturbed baseline, while R^2 decreased by 0.002300. These results indicate that changes in model forecasting performance were relatively limited under the perturbation scenarios considered in this study, although prediction errors increased slightly with increasing joint perturbation intensity. This analysis reflects the sensitivity of the model to predefined meteorological input errors and should not be directly equated with forecasting performance under actual weather forecast errors.

Based on the cross-year validation, a scenario-specific evaluation was further conducted using the independent validation set from June to August 2025 to examine the forecasting performance of the RF model under different operating conditions. The validation samples were classified into working days, weekends, extreme-heat days, normal days, high-load periods, and nighttime periods according to calendar attributes, meteorological conditions, and actual load-rate levels. High-load periods were defined as hours in which the actual load-rate fell within the highest 10% of the validation set.

For the identification of extreme-heat days, a temperature threshold of 35 °C was selected with reference to the high-temperature criteria in the *Grades of Heat Wave* (GB/T 29457—2012) [32], together with a cumulative duration requirement based on hourly meteorological data. Considering that short-term temperature peaks may not adequately represent sustained high-temperature conditions, an additional screening criterion of at least 5 h of cumulative high-temperature exposure within a day was adopted to identify operating samples under sustained high-temperature conditions. In this study, a temperature threshold of 35 °C was adopted, together with an additional cumulative duration requirement of at least 5 h per day for sample classification. The 5 h duration was defined as a scenario screening parameter in this study.

To examine the sensitivity of extreme-heat sample selection to temperature thresholds and duration requirements, three temperature thresholds (34 °C, 35 °C, and 36 °C) and three cumulative duration criteria (3 h, 5 h, and 7 h) were considered. The numbers of

extreme-heat days identified under different combinations were calculated, and the results are presented in Table 10.

Table 10. Number of Extreme-Heat Days under Different Temperature Thresholds and Cumulative Duration Criteria.

Temperature Threshold (°C)	≥3 h	≥5 h	≥7 h
≥34 °C	43	35	21
≥35 °C	34	20	10
≥36 °C	17	7	2

Note: Extreme-heat duration refers to the cumulative number of hours within a day during which the air temperature reaches or exceeds the specified temperature threshold. The values in the table indicate the number of days meeting the corresponding criteria.

As the temperature threshold and cumulative duration requirement increased, the number of extreme-heat days meeting the corresponding criteria gradually decreased. When the cumulative duration was fixed at 5 h, increasing the temperature threshold from 34 °C to 36 °C reduced the number of extreme-heat days from 35 to 7. When the temperature threshold was fixed at 35 °C, increasing the cumulative duration requirement from 3 h to 7 h reduced the number of extreme-heat days from 34 to 10. These results indicate that the number of identified extreme-heat days was sensitive to the screening criteria. Considering the high-temperature criterion, cumulative duration requirement, and sample size of the independent validation set, a temperature threshold of 35 °C with a cumulative duration of at least 5 h per day was ultimately adopted for the extreme-heat scenario evaluation.

Based on the above scenario classification method, the MAE, RMSE, MAPE, and R^2 of the RF model were calculated under different operating conditions. The results are presented in Table 11.

Table 11. Predictive Performance of the RF Model under Different Operating Conditions on the 2025 Independent Validation Set.

Operating Conditions	Number of Samples (h)	MAE (Percentage Points)	RMSE (Percentage Points)	MAPE (%)	R^2
Entire Validation Set	2040	2.0803	2.8021	3.7240	0.9296
Working Days	1440	2.1903	2.9598	3.9254	0.9252
Weekends	600	1.7832	2.3369	3.2345	0.9425
Extreme-Heat Days	480	2.5923	3.4650	4.1833	0.8975
Normal Days	1560	1.9100	2.5480	3.5803	0.9347
High-Load Periods	204	3.6646	4.6482	4.9657	−0.6889
Nighttime Periods	680	1.5995	2.1342	3.4754	0.9472

Overall, the forecasting performance of the model varied across different operating conditions. On working days, the MAE and RMSE were 2.1903 and 2.9598 percentage points, respectively, higher than the corresponding values of 1.7832 and 2.3369 percentage points on weekends. Regarding meteorological conditions, the MAPE on extreme-heat days was 4.1833%, higher than the 3.5803% observed on normal days, while R^2 decreased from 0.9347 on normal days to 0.8975 on extreme-heat days. These results indicate that the model exhibited relatively larger prediction errors under extreme-heat conditions.

In terms of load levels, the MAE and RMSE during high-load periods were 3.6646 and 4.6482 percentage points, respectively, both higher than the corresponding values for the entire validation set, with an R^2 of −0.6889. These results indicate relatively larger prediction errors in high-load samples. Meanwhile, as R^2 is affected by the range of actual

load-rate variations within this subset, it should be interpreted together with the absolute error metrics. In contrast, nighttime periods exhibited relatively lower prediction errors, with an MAE of 1.5995 percentage points, an RMSE of 2.1342 percentage points, and an R^2 of 0.9472.

3.1.4. Input Feature Ablation Experiments

To evaluate the contributions of different input features to the predictive performance of the RF model and examine the incremental predictive value of meteorological variables beyond historical load information, input feature ablation experiments were conducted while keeping the data-splitting scheme, model parameters, and training strategy unchanged. Based on the feature composition, the experiments were divided into four categories: historical load features, features without historical load information, incremental contributions of meteorological features, and the complete model. A total of 15 input feature combinations were evaluated, and the results are presented in Table 12.

As shown in Table 12, historical load features were the primary source of the model's predictive performance. When only lagged load features were used, the model achieved an MAE of 1.23 percentage points, an RMSE of 1.82 percentage points, a MAPE of 4.93%, and an R^2 of 0.9464, indicating that recent load levels and their periodic variations contained substantial predictive information. In contrast, when only rolling statistical features were used, R^2 decreased to −0.0427, indicating that the sum of squared prediction errors was slightly higher than that obtained by using the mean actual load-rate of the test set as a constant prediction. Rolling statistics primarily characterize the average levels and fluctuations within historical windows but lack a direct representation of load conditions at specific time points, which may limit their independent predictive capability. After combining lagged load and rolling statistical features, R^2 increased to 0.9510 and RMSE decreased to 1.74 percentage points, suggesting that rolling statistical features provided additional information on historical load variations beyond that captured by lagged load features.

Without historical load features, temporal and meteorological variables exhibited relatively limited predictive capability. When only temporal features were used, R^2 was −2.3024. When temperature, relative humidity, solar radiation, and wind speed were used individually, the corresponding R^2 values were −0.8961, −0.3459, −1.9929, and −1.4588, respectively. Using all original meteorological features yielded an R^2 of −0.3605. Negative R^2 values indicate that the sum of squared prediction errors exceeded the reference error obtained by using the mean actual load-rate of the corresponding test set as a constant prediction, suggesting that these features alone were insufficient to adequately characterize hourly load-rate variations at the study site. In comparison, historical load features retained temporal continuity and periodic information more directly associated with the target load-rate, providing an important basis for the model's relatively strong predictive performance. These results indicate that the independent predictive capability of different feature groups may differ from their incremental predictive value in combined models. Although rolling statistical features and temporal features yielded negative R^2 values when used independently, both provided additional predictive information when combined with historical lagged load-rate features. Therefore, the predictive value of different feature groups should be evaluated by considering both their independent forecasting performance and changes in model performance when combined with other features.

After incorporating temporal features into the complete historical load feature set, R^2 increased from 0.9510 to 0.9555, while MAE and RMSE decreased to 1.13 and 1.68 percentage points, respectively, indicating that temporal information provided additional predictive value. When individual meteorological variables were further incorporated, MAE decreased to 0.98–1.00 percentage points, RMSE decreased to 1.47–1.48 percentage

points, and R^2 increased to 0.9644–0.9651. Among the four combinations involving individual meteorological variables, the inclusion of solar radiation yielded relatively favorable results, with a MAPE of 3.95% and an R^2 of 0.9651. These results indicate that different meteorological variables provided varying degrees of incremental predictive information beyond historical load and temporal features.

Table 12. Ablation Experiment Results for Different Input Feature Combinations.

Category	Input Features	MAE (Percentage Points)	RMSE (Percentage Points)	MAPE (%)	R^2
Historical Load Features	Lagged Load Features Only	1.23	1.82	4.93	0.9464
	Rolling Statistical Features Only	6.37	8.02	30.63	−0.0427
	All Historical Load Features	1.17	1.74	4.71	0.9510
Without Historical Load Features	Temporal Features Only	13.15	14.28	58.55	−2.3024
	Temperature Only	8.21	10.82	41.27	−0.8961
	Relative Humidity Only	7.11	9.12	31.51	−0.3459
	Solar Radiation Only	11.65	13.59	59.21	−1.9929
	Wind Speed Only	10.57	12.32	53.67	−1.4588
	All Meteorological Features	6.88	9.17	29.46	−0.3605
Incremental Contribution of Meteorological Features	Historical Load + Temporal Features	1.13	1.68	4.41	0.9555
	Historical Load + Temporal + Temperature Features	1.00	1.48	4.01	0.9644
	Historical Load + Temporal + Relative Humidity Features	0.99	1.47	4.01	0.9648
	Historical Load + Temporal + Solar Radiation Features	0.98	1.47	3.95	0.9651
	Historical Load + Temporal + Wind Speed Features	0.99	1.48	4.02	0.9648
Complete Model	Historical Load + Temporal + All Meteorological Features	0.97	1.46	3.93	0.9654

When all features were included, the model achieved an MAE of 0.97 percentage points, an RMSE of 1.46 percentage points, a MAPE of 3.93%, and an R^2 of 0.9654, yielding the most favorable numerical results among the 15 input feature combinations. Compared with the model using only the complete historical load feature set, the full-feature model reduced MAE, RMSE, and MAPE by approximately 17.1%, 16.1%, and 16.6%, respectively, while increasing R^2 by 0.0144. Compared with the combination of historical load and temporal features, incorporating all meteorological features reduced RMSE from 1.68 to 1.46 percentage points and increased R^2 from 0.9555 to 0.9654.

To further examine the correlation structure among input variables, supplementary diagnostics were conducted using Pearson correlation coefficients, Spearman correlation coefficients, and the variance inflation factor (VIF). The results showed that the maximum absolute Pearson correlation coefficient among the four original meteorological variables was 0.3715, corresponding to relative humidity and wind speed. When only these variables were considered, the VIF values ranged from 1.0185 to 1.1892. After incorporating historical load features, the VIF values of the four original meteorological variables ranged from 1.0683 to 2.3201. However, some derived historical load features exhibited strong correlations. In particular, the Pearson correlation coefficient between the 24 h rolling mean and rolling maximum was 0.9764, with corresponding VIF values of 64.3954 and 64.7996, respectively. After further incorporating the squared terms of meteorological variables, the VIF values for the original temperature and relative humidity terms increased to 39.3393

and 47.3227, respectively, indicating that nonlinear feature expansion may increase linear redundancy among input variables. These findings suggest that the predictive contributions of meteorological variables should be interpreted in conjunction with their correlations with other input features. The ablation results reflect changes in predictive performance under specific feature combinations and should not be directly interpreted as independent causal effects of individual variables on actual transformer load-rate.

Overall, historical load features constituted the primary basis for the predictive performance of the RF model, while temporal and meteorological features provided additional predictive information to varying degrees. The inclusion of meteorological variables improved predictive performance on the test set examined in this study, although their incremental contributions were associated with the presence of other input features and the correlation structure among them. These findings provide a basis for subsequent SHAP and ALE analyses of the attribution contributions and nonlinear response characteristics of meteorological variables in the model.

3.2. Impact of Meteorological Parameters on the Load-Rate

3.2.1. Meteorological Feature Contribution Analysis

To analyze the attribution contributions of meteorological features to load-rate predictions, the mean absolute SHAP values of individual meteorological features were calculated across the evaluation samples using the trained RF model. Figure 7 presents the relative importance of the meteorological features. The bar length represents the mean absolute SHAP value, with larger values indicating greater average attribution magnitudes under the current model and evaluation samples.

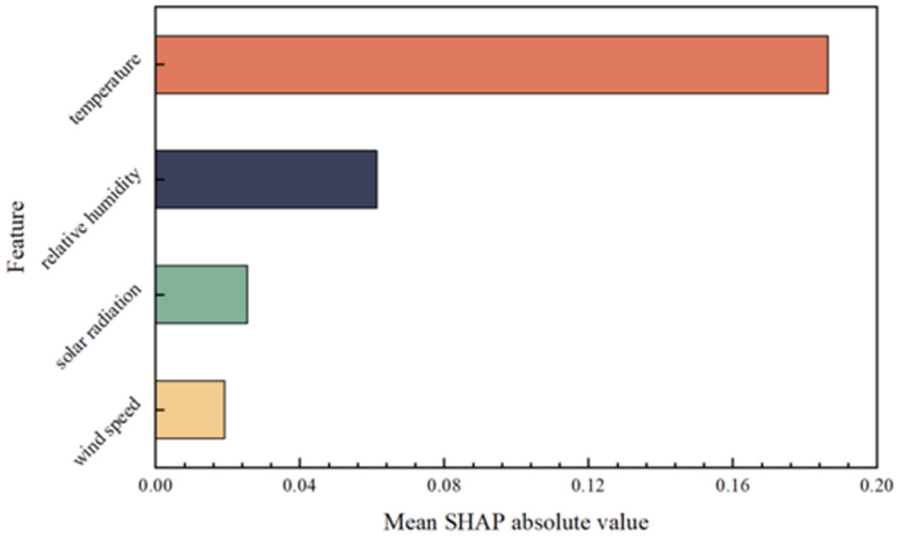


Figure 7. Average Absolute SHAP Values for Each Meteorological Parameter.

As shown in Figure 7, temperature had the highest mean absolute SHAP value of 0.1864, followed by relative humidity at 0.0613, while solar radiation and wind speed exhibited values of 0.0254 and 0.0191, respectively. The mean absolute SHAP value of temperature was approximately 3.04 times that of relative humidity, indicating that temperature made a relatively greater attribution contribution to model predictions under the current RF model and feature system. Overall, the four meteorological variables were ranked in descending order of attribution importance as temperature, relative humidity, solar radiation, and wind speed, indicating differences in their average attribution magnitudes to model predictions.

To further evaluate the statistical stability of meteorological feature attributions, Bootstrap resampling was employed to calculate the 95% empirical intervals of the mean absolute SHAP values for each variable. The results are presented in Table 13.

Table 13. Bootstrap Stability Analysis of Meteorological Feature Attributions.

Meteorological Variable	Mean |SHAP|	95% Bootstrap CI
Temperature	0.1864	[0.1805, 0.1933]
Relative humidity	0.0613	[0.0576, 0.0651]
Solar radiation	0.0254	[0.0223, 0.0287]
Wind speed	0.0191	[0.0174, 0.0209]

Note: The 95% bootstrap confidence intervals were obtained from 1000 resamples using the fixed RF model. The intervals represent the sampling variability of mean absolute SHAP values under the specified resampling procedure.

The 95% empirical intervals were [0.1805, 0.1933] for temperature, [0.0576, 0.0651] for relative humidity, [0.0223, 0.0287] for solar radiation, and [0.0174, 0.0209] for wind speed. The empirical intervals of all four meteorological variables were relatively narrow and showed no overlap, indicating relatively clear differences in their mean attribution contributions under the Bootstrap resampling conditions of the current RF model and evaluation samples.

Combined with the feature ablation experiments and correlation diagnostics, these findings indicate that meteorological variables provided a certain degree of incremental predictive information beyond historical load and temporal features. However, their feature attributions may still be affected by correlations among input variables and redundancy in derived features. Therefore, the Bootstrap results primarily reflect the stability of feature attributions under variations in the evaluation samples using the fixed RF model. These findings do not account for attribution stability arising from model retraining, nor do they establish independent causal effects of meteorological variables on actual transformer load-rate.

3.2.2. Climate Feature Response Analysis

To further analyze the relationships between the values of meteorological variables and their attribution contributions to model predictions, SHAP summary plots were generated for the four meteorological features, as shown in Figure 8. In the figure, each point represents the SHAP value of an individual sample. The horizontal axis indicates the attribution contribution of the feature to the model prediction and its direction, while the color represents the corresponding feature value level.

As shown in Figure 8, the SHAP values for temperature exhibited a relatively wide distribution and generally shifted from negative to positive contributions as temperature increased. Samples with lower temperatures were mostly associated with negative SHAP values, whereas those with higher temperatures were mainly concentrated in the positive SHAP region, indicating that the model tended to produce relatively higher load-rate predictions under higher-temperature conditions. Relative humidity also exhibited a certain degree of value-dependent variation, with SHAP values for high-humidity samples generally concentrated in the positive region, while those for low-humidity samples were more frequently distributed in the negative region. In contrast, the SHAP values for solar radiation and wind speed were more concentrated, with relatively smaller overall attribution magnitudes. Solar radiation exhibited both positive and negative contributions across different feature values, whereas most SHAP values for wind speed were distributed near zero.

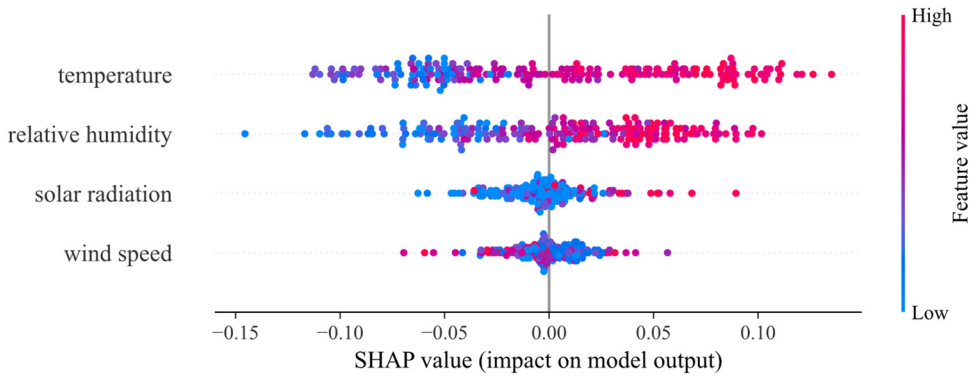


Figure 8. SHAP Distributions for Each Meteorological Parameter. *Note: The vertical reference line indicates a SHAP value of zero.*

Considering the potential correlations among different meteorological parameters, SHAP dependence plots were further used to examine how the contribution of temperature varied under different relative humidity conditions (Figure 9). In the figure, the horizontal axis represents temperature, the vertical axis represents the SHAP value of the temperature feature, and the color indicates the level of relative humidity.

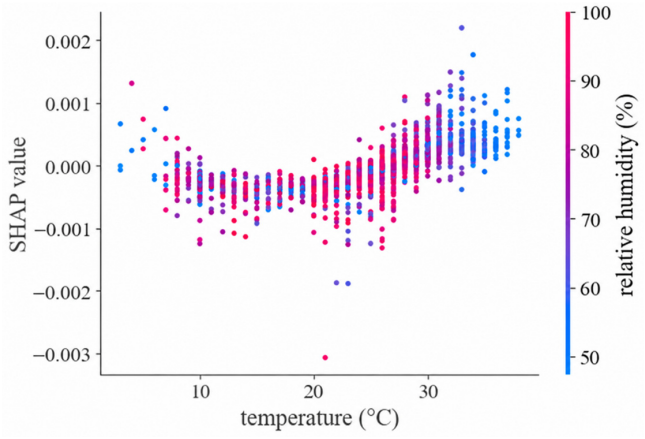


Figure 9. SHAP Dependence Plot for Temperature, with Color Indicating Relative Humidity.

As shown in the figure, within the temperature range of approximately 10–20 °C, the SHAP values for temperature were distributed mainly in the negative region, indicating that this temperature range generally made a negative contribution to the model predictions. As temperature increased further, the SHAP values gradually increased and were predominantly positive at higher temperatures. The distributions of temperature SHAP values differed to some extent across relative humidity levels, although considerable overlap remained among the humidity groups, suggesting that the contribution of temperature to the model output may vary with relative humidity conditions.

3.2.3. ALE Analysis

To further analyze the nonlinear response characteristics of the RF model to meteorological variables, ALE curves were generated for temperature, relative humidity, solar radiation, and wind speed, as shown in Figure 10. The horizontal axis represents the values of the meteorological variables, while the vertical axis represents the centered ALE values in percentage points. By comparing the shapes of the curves for different meteorological variables, the average local response trends of model predictions associated with variations in meteorological conditions were examined.

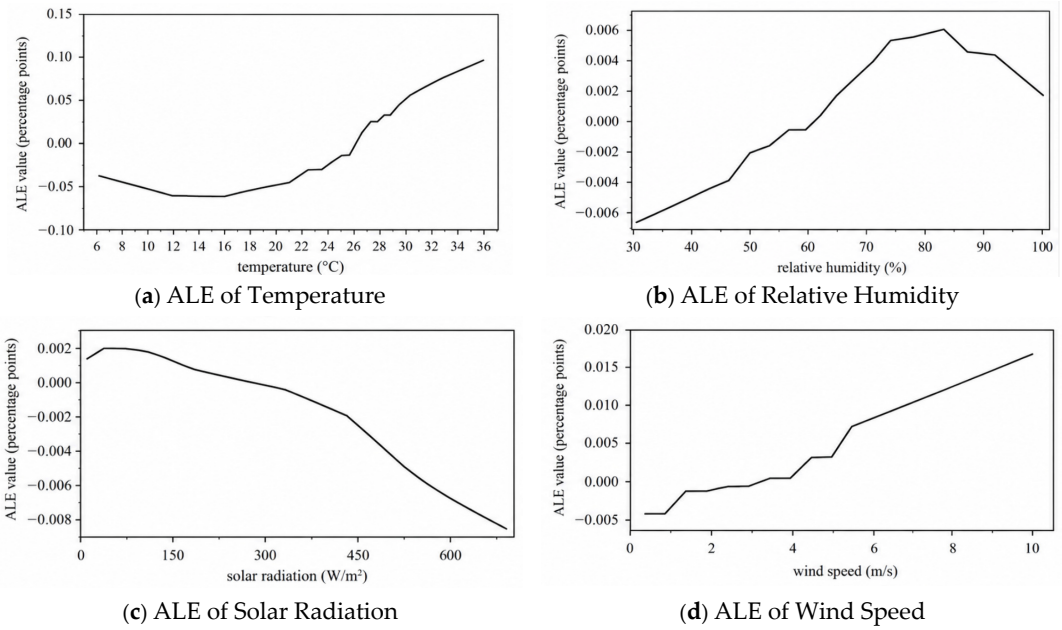


Figure 10. ALE Curves Between Meteorological Parameters and RF Model Predictions.

Figure 10a shows that the temperature ALE value decreased slightly in the low-temperature range and remained relatively stable at approximately 12–18 °C. It then increased gradually above 20 °C, with a more rapid rise observed within approximately 24–30 °C. Overall, the model output exhibited a pattern of relative stability followed by an increase as temperature rose.

Figure 10b shows that, under the baseline configuration of 20 equal-frequency bins, the ALE values for relative humidity exhibited a nonlinear pattern characterized by an initial increase followed by a decline as humidity increased. Specifically, the ALE values gradually increased in the low-to-moderate humidity range, reached a relatively high level at higher humidity levels, and subsequently showed a downward trend. Considering that the local shape of the relative humidity ALE curve was somewhat sensitive to the binning configuration, these results primarily reflect the model-predicted response characteristics under the baseline binning configuration, while the specific locations of the turning points remain subject to some uncertainty.

Figure 10c shows that, under the baseline configuration of 20 equal-frequency bins, the ALE curve for solar radiation exhibited an overall downward trend, with some differences in the magnitude of local variations across different radiation ranges. Combined with the binning sensitivity analysis results, the local response patterns of solar radiation exhibited some variability with changes in the number of bins, indicating a certain degree of binning dependence in its nonlinear response characteristics. Therefore, these results primarily characterize the overall association between solar radiation and model predictions under the current RF model, while local turning points should not be interpreted as response thresholds with clearly established physical significance.

Figure 10d shows that the wind speed ALE value generally increased with increasing wind speed. The curve changed only slightly in the low-wind-speed range, exhibited a stepwise increase at approximately 4–6 m/s, and continued to rise above 6 m/s. Compared with the other meteorological parameters, the overall variation in the wind speed ALE value was relatively small.

Overall, the ALE curves for temperature, relative humidity, solar radiation, and wind speed exhibited varying degrees of nonlinearity. Temperature showed a relatively pronounced upward trend at higher values; relative humidity displayed an increase followed

by a decline; solar radiation generally showed a downward trend, with a more rapid decrease in the high-radiation range; and wind speed exhibited a gradual overall increase.

To examine the sensitivity of ALE curves to the number of bins, 20 equal-frequency bins were used as the baseline, with additional comparisons conducted using 10 and 30 bins. The results are presented in Table 14.

Table 14. Correlations between ALE curves under different bin settings.

Meteorological Variable	10 vs. 20 bins (r)	30 vs. 20 bins (r)
Temperature	0.9974	0.9996
Relative humidity	0.7351	0.9283
Solar radiation	0.8450	0.9523
Wind speed	0.9988	1.0000

As shown in Table 14, the ALE curves for different meteorological variables exhibited varying degrees of sensitivity to the number of bins. For temperature and wind speed, the Pearson correlation coefficients between the curves obtained using 10, 20, and 30 bins all exceeded 0.99, indicating that their overall model-predicted response trends were relatively insensitive to the selected number of bins. In contrast, the ALE curves for relative humidity and solar radiation showed lower consistency across different binning configurations. Specifically, the Pearson correlation coefficients between the 10-bin and 20-bin curves were 0.7351 and 0.8450 for relative humidity and solar radiation, respectively, increasing to 0.9283 and 0.9523 when comparing the 30-bin and 20-bin curves. These results indicate that the response curves of these two variables exhibited a certain degree of binning dependence, with potential differences in their local shapes under different binning configurations. Overall, ALE analysis can characterize the nonlinear associations between meteorological variables and model predictions under the current RF model. However, the local turning points and response magnitudes for relative humidity and solar radiation should be interpreted with caution and should not be directly regarded as stable physical response thresholds.

Considering both the ALE response curves and binning sensitivity analysis results, the overall response trends for temperature and wind speed showed relatively high consistency across different binning configurations, whereas the local response patterns of relative humidity and solar radiation were more sensitive to the number of bins. These findings characterize the nonlinear associations of different meteorological variables with predictions under the current RF model and the corresponding scope of interpretation.

4. Conclusions

Using the Panzhou regional substation in Maoming City, Guangdong Province, China, as the research case, this study evaluated the forecasting performance of the RF model and conducted input feature ablation experiments and meteorological feature interpretability analyses based on hourly transformer load-rate and observed meteorological data. The study focused on distinguishing the predictive contributions of historical load and meteorological information and examining the associations between meteorological variables and model outputs. The results indicate that historical load information constituted the primary basis for the model's relatively high forecasting accuracy, while meteorological variables provided additional predictive information beyond historical load features. Through model validation under different temporal conditions and SHAP and ALE analyses, this study provides case-based evidence for transformer load-rate forecasting and meteorological feature interpretation at regional substations. The main conclusions are as follows:

(1) Under the condition that observed meteorological data at the target time were available, the models exhibited differences in forecasting performance for hourly transformer load-rate prediction at regional substations. On the independent test set for 2024, LightGBM achieved an MAE of 0.915 percentage points, an RMSE of 1.337 percentage points, a MAPE of 3.674%, and an R^2 of 0.9710, yielding relatively favorable forecasting results among the comparison models. The corresponding metrics for RF were 0.97 percentage points, 1.46 percentage points, 3.93%, and 0.9654, respectively, indicating relatively strong predictive performance. Rolling-origin validation and statistical significance tests showed that LightGBM exhibited statistically significant performance advantages over RF in terms of MAE, RMSE, and MAPE, while RF outperformed LSTM and XGBoost in terms of MAE and MAPE, with statistically significant differences. Although forecasting performance varied among the models, RF maintained relatively strong overall predictive performance under the dataset and evaluation conditions considered in this study.

(2) On the cross-year independent validation set from June to August 2025 at the same substation, the RF model achieved an R^2 of 0.9296, a MAPE of 3.72%, an MAE of 2.08 percentage points, and an RMSE of 2.80 percentage points, indicating that it retained a certain degree of predictive capability on cross-year warm-season data from the same site. The scenario-specific evaluation showed that the MAPE and R^2 were 4.18% and 0.8975, respectively, on extreme-heat days, compared with 3.58% and 0.9347 on normal days. During high-load periods, the MAE and RMSE increased to 3.66 and 4.65 percentage points, respectively, indicating relatively larger prediction errors under extreme-heat and high-load conditions. The meteorological input perturbation experiment showed that changes in model performance were relatively limited within the perturbation ranges considered in this study. Under the high-intensity joint perturbation scenario, RMSE increased by approximately 1.21% compared with the unperturbed baseline. These findings reflect the model's cross-year forecasting performance under warm-season conditions at the same study site and its sensitivity to simulated meteorological input errors.

(3) The ablation experiments showed that historical load features constituted the primary information basis for the predictive performance of the RF model, while temporal and meteorological features provided varying degrees of incremental predictive information. When all historical load features were used alone, the model achieved an R^2 of 0.9510. After further incorporating temporal and meteorological features, R^2 increased to 0.9654, while RMSE decreased by approximately 16.1%. Taking the combination of historical load and temporal features as the baseline, the individual inclusion of temperature, relative humidity, solar radiation, and wind speed increased R^2 by 0.0089, 0.0093, 0.0096, and 0.0093, respectively. These results indicate that each meteorological parameter provided some additional predictive value, although its incremental contribution was relatively limited.

(4) Under the complete feature system adopted at the study site, the mean absolute SHAP values for temperature, relative humidity, solar radiation, and wind speed were 0.1864, 0.0613, 0.0254, and 0.0191, respectively, with temperature and relative humidity exhibiting relatively higher attribution contributions to model predictions. The Bootstrap resampling analysis showed that the 95% empirical intervals of the mean absolute SHAP values for the meteorological variables were relatively narrow and did not overlap, indicating a certain degree of stability in their relative attribution contributions under the current resampling conditions. The ALE analysis showed that the model tended to produce higher load-rate predictions under higher-temperature conditions. Relative humidity exhibited a nonlinear response characterized by an initial

increase followed by a decline, while solar radiation showed an overall downward trend and wind speed exhibited relatively gradual variations. The binning sensitivity analysis indicated that the overall response trends for temperature and wind speed were highly consistent across different binning configurations, whereas the local response patterns of relative humidity and solar radiation were more sensitive to the number of bins. These findings characterize the nonlinear associations between meteorological variables and model predictions under the current model, while the specific locations of local turning points should be interpreted with caution.

The limitations of this study are primarily associated with the scope of the dataset, the form of meteorological inputs, and the level of practical application. The study was based on operational data from a 40 MVA main transformer at the Panzhou regional substation in Maoming City, Guangdong Province, China. Data from 2024 were used for model development, rolling-origin validation, and seasonal performance evaluation, while cross-year validation was limited to warm-season samples from June to August 2025 at the same site. Therefore, the findings primarily reflect the model's forecasting performance under different temporal conditions at the study site, and its applicability to other substations and other seasons across different years requires further validation. Model training, validation, and interpretability analyses were all conducted using observed meteorological data at the target time. Although predefined random perturbations were introduced to evaluate the sensitivity of model performance to meteorological input errors, this experiment cannot replace an analysis of error propagation under actual weather forecast conditions. Forecasting performance associated with different forecast lead times and weather forecast products remains to be further investigated. In addition, the SHAP and ALE results reflect feature attribution contributions and model-predicted responses under the given model, feature system, and sample distribution. Their interpretations may be affected by feature correlations, redundancy in derived features, and ALE binning configurations, and do not establish independent causal relationships between meteorological variables and actual transformer load-rate. The point predictions, prediction intervals, and threshold exceedance probabilities obtained in this study can provide information on load levels and prediction uncertainty. However, the quantitative relationships between prediction errors and load allocation, dispatch strategies, and economic costs require further investigation in conjunction with actual equipment constraints, power supply reliability requirements, and operational cost models.

Author Contributions: Conceptualization, Z.C. and Y.D.; methodology, Z.C., Y.W., Y.D. and H.W.; software, Y.W.; validation, Y.W.; formal analysis, Z.C. and Y.W.; investigation, L.L.; resources, L.L.; data curation, L.L. and Y.W.; writing—original draft preparation, Z.C. and Y.W.; writing—review and editing, Y.D.; supervision, H.W.; project administration, Y.D.; funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the First Batch of Infrastructure Project Special Service Projects of Guangdong Power Grid Co., Ltd. in 2024 (Grant No. 0309002025030103JJ00003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Authors Zhongmin Chen and Lei Li were employed by Maoming Power Supply Bureau of Guangdong Grid Co., Ltd. The remaining authors declare that the research was conducted in the Absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Abbreviations

The following nomenclature and abbreviations are used in this manuscript:

Nomenclature

S_{actual}	the apparent power at time t
S_{rate}	the rated capacity of the transformer
φ_i	the SHAP value for feature i
N	the set of all features
S	the subset of features excluding feature i
n	the total number of elements in set N
x_s	the target feature set
x_c	the complement of the set

Abbreviations

SCADA	Supervisory Control and Data Acquisition
PRECON	Pakistan Residential Electricity Consumption
SARIMA	Seasonal Autoregressive Integrated Moving Average
IEA	International Energy Agency
SVR	Support Vector Regression algorithm
CMDC	China Meteorological Data Service Centre
XGBoost	Extreme Gradient Boosting algorithm
LSTM	Long Short-term Memory
RCP	Representative Concentration Pathways
ADF	Augmented Dickey–Fuller
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
VIF	Variance Inflation Factor
SHAP	Shapley Additive Explanations
ALE	Accumulated Local Effects
PDP	Partial Dependence Plot
RF	Random Forest
R^2	Coefficient of Determination
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MBE	Mean Bias Error
MaxE	Maximum Error

References

1. International Energy Agency. *Electricity 2024*; International Energy Agency: Paris, France, 2024. Available online: https://www.iea.org/reports/electricity-2024 (accessed on 1 March 2026).
2. Huang, Z.; Li, X.; Lu, B. Saturated Electricity Demand Forecasting under the "Dual Carbon" Goals: Characteristics, Criteria, and Multi-Method Comparisons. *Hongshui River* **2025**, *3*, 8–14. [CrossRef]
3. Rizi, E.T.; Rastegar, M.; Forootani, A. Power system flexibility analysis using net-load forecasting based on deep learning considering distributed energy sources and electric vehicles. *Comput. Electr. Eng.* **2024**, *117*, 109305. [CrossRef]
4. Ma, L.; Li, Y. A Characteristic Extraction and Forecast Method for Interval Power Loads. *Control Eng. China* **2015**, *22*, 645–648.
5. Ullah, S.; Liguo, W.; Amin, S.U.; Alam Khan, A.; Dreglea, A.; Sami, I.; Sidorov, D. Toward robust household load forecasting in smart grids using uncertainty-aware and cross-regional hybrid deep learning. *Comput. Electr. Eng.* **2026**, *138*, 111319. [CrossRef]
6. International Energy Agency. *Global Energy Review 2025*; International Energy Agency: Paris, France, 2025. Available online: https://www.iea.org/reports/global-energy-review-2025 (accessed on 1 March 2026).
7. Yang, Z.; Li, W.; Yan, Z. The Relationship between Temperature Change and Electric Power Demand—Evidence from China's Urban Panel Data from 2000 to 2014. *J. Beijing Inst. Technol.* **2019**, *5*, 44–55. [CrossRef]
8. Bashmakov, I.A.; Nilsson, L.J.; Acquaye, A.; Bataille, C.; Cullen, J.M.; Fischedick, M.; Geng, Y.; Tanaka, K. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2022.
9. U.S. Energy Information Administration. *Annual Energy Outlook 2023*; U.S. Energy Information Administration: Washington, DC, USA, 2023. Available online: https://www.eia.gov/outlooks/aeo/archive.php (accessed on 1 March 2026).

10. Chen, G.; Lu, S.; Zhou, S.; Tian, Z.; Kim, M.K.; Liu, J.; Liu, X. A Systematic Review of Building Energy Consumption Prediction: From Perspectives of Load Classification, Data-Driven Frameworks, and Future Directions. *Appl. Sci.* **2025**, *15*, 3086. [CrossRef]
11. Lee, H.-H.; Arthur, R.S.; Golaz, J.-C.; Edmunds, T.A.; Wert, J.L.; Signorotti, M.V.; Watson, J.-P. Assessment of Climate Change Impacts on Renewable Energy Resources in Western North America. *Energies* **2025**, *18*, 3467. [CrossRef]
12. Wang, T.; Su, X.; Mei, Y.; Xiong, W. Climate change intensifies low-carbon transition pressure in China's power system. *Environ. Impact Assess. Rev.* **2025**, *114*, 107949. [CrossRef]
13. Chen, J.Q.; Yang, J.J.; Lou, Z.B. A Novel Short-Term Load Forecasting Model Based on XGBoost Algorithm. *Electr. Meas. Instrum.* **2019**, *56*, 23–29. [CrossRef]
14. Deng, Y.X.; Huang, Y.P.; Huang, Z.C. Electric Vehicle Charge-Discharge Capacity Prediction Based on Random Forest Algorithm. *Autom. Electr. Power Syst.* **2021**, *45*, 181–188. [CrossRef]
15. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2024; Volume 37. [CrossRef]
16. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
17. Guo, Q.; Tian, Z.; Ding, Y.; Zhu, N. An improved office building cooling load prediction model based on multivariable linear regression. *Energy Build.* **2015**, *107*, 445–455. [CrossRef]
18. Wang, R.; Lu, S.; Feng, W. A novel improved model for building energy consumption prediction based on model integration. *Appl. Energy* **2020**, *262*, 114561. [CrossRef]
19. Wu, W.; Deng, Q.; Shan, X.; Miao, L.; Wang, R.; Ren, Z. Short-Term Forecasting of Daily Electricity of Different Campus Building Clusters Based on a Combined Forecasting Model. *Buildings* **2023**, *13*, 2721. [CrossRef]
20. Zhang, G.H.; Li, J.X.; Zhang, X.C.; Qin, Y.Y.; Jin, X.L. Coordinated Optimization of Buildings and Integrated Community Energy Systems Through a Bi-level Model Predictive Control Method with Stackelberg Game. *J. Glob. Energy Interconnect.* **2025**, *8*, 289–298. [CrossRef]
21. State Grid Henan Electric Power Co., Ltd.; Electric Power Research Institute of State Grid Henan Electric Power Co., Ltd. A Random Forest-Based Method for Predicting Fault Risk Levels in Distribution Networks under Persistent High-Temperature Conditions. Chinese Patent CN116663393A, 29 August 2023. Available online: https://patents.google.com/patent/CN116663393A/zh (accessed on 1 March 2026).
22. Chen, X.; Zou, Q.N.; Xie, S.Y.; Chen, C.Q. State recognition of pressure plate switches based on transfer learning strategy. *Comput. Mod.* **2021**, *5*, 120–126. [CrossRef]
23. Zohaib, A.; Akram, F.; Khalid, S.; Nawaz, H.; Rehman, M.U. Hybrid deep learning based load forecasting and AI-driven energy management for grid-connected multi-microgrids. *Comput. Electr. Eng.* **2026**, *131*, 110915. [CrossRef]
24. Wu, K.; Gu, J.; Meng, L.; Wen, H.; Ma, J. An explainable framework for load forecasting of a regional integrated energy system based on coupled features and multi-task learning. *Prot. Control Mod. Power Syst.* **2022**, *7*, 24. [CrossRef]
25. Amin, N.; Kim, Y.-W.; Byun, Y.-C. ProphetBoost: A hybrid pipeline for accurate and transparent electricity load forecasting. *Comput. Electr. Eng.* **2026**, *138*, 111280. [CrossRef]
26. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
27. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
28. Said, S.E.; Dickey, D.A. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **1984**, *71*, 599–607. [CrossRef]
29. Kwiatkowski, D.; Phillips, P.C.B.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econom.* **1992**, *54*, 159–178. [CrossRef]
30. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
31. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4768–4777.
32. *GB/T 29457—2012*; General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China; Standardization Administration of China. Grade of the Heat Wave. China Standards Press: Beijing, China, 2012.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.