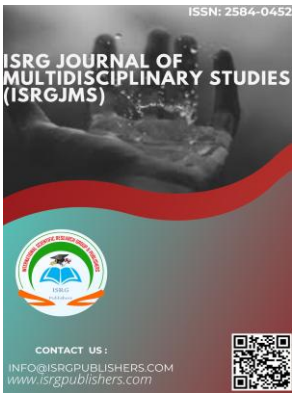
ISRG Journal of Multidisciplinary Studies (ISRGJMS)

ISRG PUBLISHERS
Abbreviated Key Title: isrg j. multidiscip. Stud.
ISSN: 2584-0452 (Online)
Journal homepage: https://isrgpublishers.com/isrgjms/
Volume – IV, Issue - IV (April) 2026
Frequency: Monthly



Electricity Demand Forecasting Using Transformer-Based Model Optimization and Interval Calibration

Pelin Li[1] Qingqing Hao[2] Xingan Li[3*]

[1] Faculty of Law, Peking University, China.

[2] Department of Data Science, Durham University, United Kingdom.

[3] Beihang University, China.

| **Received:** 15.03.2026 | **Accepted:** 19.03.2026 | **Published:** 14.04.2026

***Corresponding author:** Xingan Li

Abstract

This study proposes a power load prediction framework based on multiple machine learning models, including Naive 7, XGBoost, Random Forest, and a deep learning model based on Transformer. The framework integrates hyperparameter optimization, time-series cross-validation, and uncertainty quantification using the conformal prediction (CP) method. The results show that the Transformer model achieves higher accuracy than the other models, with the lowest MAE, RMSE, and MAPE, and the highest R^2 and PICP. The Transformer model is effective at capturing long-term dependencies and providing a reliable forecast range, making it suitable for dynamic power systems.

Keywords: *Transformer Model; Electricity Load Forecasting; Machine Learning; Conformal Prediction; Hyperparameter Optimization; Uncertainty Quantification; Smart Grid Forecasting*

1. Introduction

1.1 Motivation

Energy demand forecasting plays a vital role in ensuring the efficient operation of power systems and in optimizing energy production and distribution. Accurate forecasting helps grid operators, utilities, and policymakers plan for future energy demand, prevent electricity shortages, and minimize operating costs. However, technological progress, socioeconomic changes, and factors such as government policy have led to increasingly complex energy systems, which pose significant challenges to traditional forecasting methods. In addition, the rapid growth of renewable energy sources, such as wind and solar power, introduces further uncertainty due to their inherent variability. As a result, more advanced and reliable forecasting techniques are

DOI: 10.5281/zenodo.19563630

required to capture both short-term and long-term fluctuations in electricity demand.

In recent years, machine learning approaches, including both classical methods and deep learning models, have gained prominence due to their ability to capture complex nonlinear relationships in energy consumption data. Tree-based models such as XGBoost and Random Forest are widely applied because of their flexibility in handling nonlinearities and interactions among features. Meanwhile, deep learning architectures such as the Transformer have demonstrated strong potential in modeling long-term dependencies and complex temporal patterns in time-series data. Furthermore, recent developments in uncertainty quantification, particularly through conformal prediction (CP) intervals, provide a systematic way to evaluate the reliability of forecasts and offer valuable insights into potential variations in energy demand.

This study aims to develop and evaluate a comprehensive framework for power load forecasting that integrates multiple machine learning models, hyperparameter optimization, time-series cross-validation, and uncertainty quantification using conformal prediction. The models considered in this study include the Naive 7 benchmark model, XGBoost, Random Forest, and a Transformer-based deep learning model. The primary objective is to assess the accuracy, reliability, and uncertainty of the forecasting results, while also examining the influence of key features within the prediction models. Through this work, the study seeks to contribute to the development of more reliable and adaptive forecasting approaches capable of supporting modern energy systems that are increasingly dynamic and uncertain.

2. Literature Review

The emerging complexities caused by economic development, technological progress, and government policy, together with the requirement for low-carbon development of power grids, have created significant challenges for power system coordination and operation (Jiang et al., 2020). To address these challenges, intelligent and optimization-based forecasting models have been widely applied. For example, the optimized number of neurons in the hidden layers of a multi-layer perceptron (MLP) can be determined using the particle swarm optimization (PSO) algorithm, which effectively improves the prediction accuracy of time-series load data (Sheikhan & Mohammadi, 2013). Economic development is often associated with increased energy consumption in industrial and urban sectors, suggesting that demand forecasting models should account for socioeconomic dynamics in addition to technical parameters (Tilahun, Bhandari, & Mamo, 2019). In this context, hybrid models have also been investigated to address prediction interval and density estimation problems and have become increasingly common in short-term energy consumption forecasting (Khairalla et al., 2018).

In statistical modeling traditions, ambiguity in model calibration is typically interpreted as over-parameterization, which introduces uncertainty into the forecasting process (van Ruijven et al., 2010). Demand forecasting plays a vital role in energy supply–demand management for both governments and private companies. Therefore, the development of models capable of accurately forecasting future energy consumption trends—particularly for nonlinear data—has become an important issue in power production and distribution systems (Ghalehkhondabi et al., 2017). However, when strong nonlinearities exist in the forecasting problem, linear approaches may fail to capture the underlying dynamics of the process. In such cases, hybrid neuro-fuzzy models have been proposed as an effective alternative for mid-term energy demand forecasting (Iranmanesh, Abdollahzade, & Miranian, 2012).

3. Methodology

In this study, we employ a comprehensive electricity load forecasting framework that integrates multiple machine learning models, hyperparameter optimization (HPO), time-series cross-validation (CV), and forecast intervals calibrated using conformal prediction (CP). The specific methods are as follows.

3.1 Model Selection and Construction

We compare several classical and advanced models.

Naive-7 model: As a benchmark, this model predicts future load based on the average load of the previous seven days. Although simple, it provides a reference point for evaluating improvements achieved by more complex models.

Tree-based models (XGBoost and Random Forest): These models were selected for their ability to handle nonlinear relationships and interactions between features. Bayesian optimization (Optuna) is used to fine-tune the hyperparameters of these models to improve forecasting performance.

Deep learning model based on Transformer: To capture long-term dependencies, a Transformer-based architecture is designed specifically for time-series prediction. The model combines Fourier transforms, lag features, and rolling statistics to effectively capture seasonal and temporal patterns. The Transformer model is trained using the Huber loss function and gradient clipping (clipnorm) to ensure stable convergence.

In addition, the Transformer model generates forecast intervals using conformal prediction (CP). The CP approach is combined with validation-driven calibration, drift-aware safety margins, and Mondrian group scaling (adjusting the interval width monthly) to optimize interval accuracy.

3.2 Data Preprocessing and Feature Engineering

The raw dataset contains electricity load data for England and Wales from 2014 to 2024. Various exogenous variables are incorporated, including temperature, precipitation, solar radiation, and renewable energy generation data (such as wind and photovoltaic power generation). These external variables help capture both demand-side and supply-side factors that influence load forecasting.

The preprocessing of raw data follows a strict procedure to ensure data quality and consistency. Missing values are handled using linear interpolation, which maintains the temporal continuity of the series while avoiding the introduction of significant bias. Outliers in load and weather data are identified and corrected using a combination of z-score analysis and domain-specific rules. In addition, additional features are extracted from calendar information, such as working days, months, and public holidays, to capture underlying seasonal and weekly patterns.

3.3 Feature Engineering for Improved Model Performance

To enhance the model's ability to capture temporal dependencies, a set of lag features is constructed using historical power load values. The specific lag periods include 1, 7, 14, and 28 days. These lag

DOI: 10.5281/zenodo.19563630

features enable the model to capture both short-term and long-term autocorrelations in the data.

Rolling statistics are also computed, including moving averages and standard deviations over 7-day, 14-day, and 28-day windows. These rolling statistics help smooth recent trends and fluctuations, which is particularly useful for models sensitive to abrupt changes.

To capture the inherent seasonal patterns of electricity consumption, Fourier series decomposition is applied to generate sine and cosine components for daily, weekly, and yearly periods. This transformation explicitly models cyclical changes rather than relying solely on historical load values.

In addition, categorical features (e.g., date types such as weekdays, weekends, and holidays) are encoded using one-hot encoding to provide clear representations for both tree-based models and neural networks.

3.4 Model Evaluation and Metrics

Model performance is evaluated using several complementary metrics. The accuracy of point predictions is measured using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). Prediction intervals are evaluated using prediction interval coverage probability (PICP), which measures the proportion of actual observations captured within the constructed interval, and mean interval width (MIW), which quantifies the sharpness of the interval. These indicators provide a comprehensive evaluation of both accuracy and reliability.

In addition to these performance metrics, conformal prediction methods are used to estimate prediction intervals. This approach ensures reliable coverage even with relatively small samples and allows the quantification of predictive uncertainty. For the Transformer model, residuals are normalized using recent rolling standard deviations, and the intervals are seasonally adjusted to reflect expected volatility. This provides decision-makers with practical insights into expected demand and its associated uncertainty.

3.5 Model Training and Hyperparameter Optimization

All models are trained using a rolling-window approach based on historical data to ensure that no future information is used during training. The dataset is divided chronologically into training and testing sets, with the most recent two years reserved for testing.

Hyperparameter optimization for the tree-based models involves tuning parameters such as the number of trees, tree depth, learning rate, sampling rate, and estimator parameters. For the Transformer model, the training process includes the use of the Adam optimizer with a learning-rate scheduler, early stopping, and gradient clipping to ensure stable convergence.

Loss functions are selected to balance robustness and sensitivity to extreme errors. Tree-based models are optimized using mean squared error (MSE) or mean absolute error (MAE), while the Transformer model employs the Huber loss function to mitigate the impact of outliers, which commonly occur during extreme weather events or holiday periods in power load data.

4. Results

4.1 Overall predictive performance

Table 1. Metrics Comparison

Model	MAE	RMSE	MAPE	R^2
Naive-7	1559.08	2059.96	6.37	0.706
XGBoost	1206.21	1569.61	5.08	0.829
RandomForest	1286.64	1696.26	5.38	0.800
Transformer	1026.30	1342.32	4.25	0.875

Figure 1. Metrics Comparison

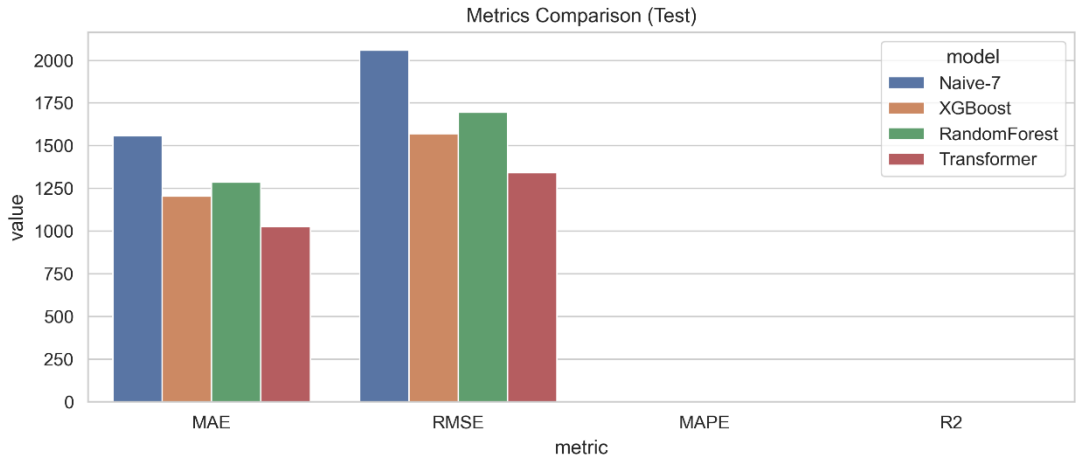


Table 1 and Figure 1 present a comparison of the four models (Naive-7, XGBoost, Random Forest, and Transformer) based on several performance indicators. The Transformer model achieves the lowest MAE, at 1026.30, followed by XGBoost (1206.21), Random Forest (1286.64), and Naive-7 (1559.08).

DOI: 10.5281/zenodo.19563630

Regarding RMSE (root mean square error) and MAPE (mean absolute percentage error), the Transformer model also performs best, with the lowest RMSE and MAPE, at 1342.32 and 4.25%, respectively. In addition, the Transformer model achieves the highest R-squared value of 0.875, indicating the best fit to the actual demand data. This is followed by XGBoost (0.829), Random Forest (0.800), and Naive-7 (0.706).

4.2 Time series fitting and temporal tracking

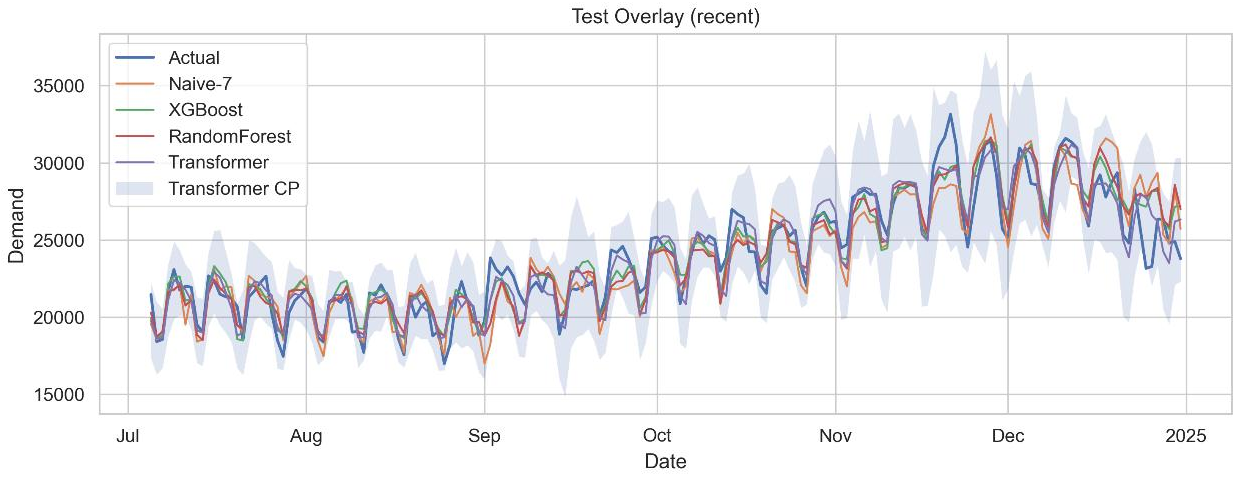


Figure 2. Test Overlay (recent)

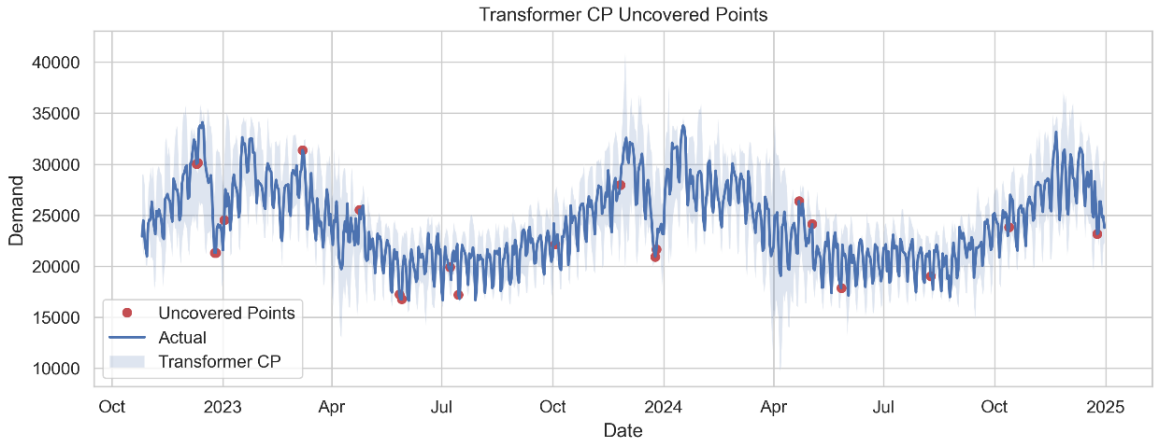


Figure 3. Transformer CP Uncovered Points

Figure 2 shows the prediction results for actual demand and the forecasts generated by the Naive-7, XGBoost, Random Forest, and Transformer models, together with the prediction intervals produced by the Transformer model using conformal prediction (CP). The actual demand is shown in blue, while the forecast results for each model are displayed in different colors. The shaded area represents the confidence interval generated by the Transformer CP.

Figure 3 illustrates the actual power demand over time (from October 2023 to October 2025), where the prediction interval of the Transformer model is shaded in light blue. Red dots indicate points where the actual demand falls outside the forecast interval; these are referred to as "uncovered points." These uncovered points represent situations in which the model forecasts fail to capture the actual demand, highlighting periods of higher uncertainty or volatility in the demand data.

4.3 Prediction–observation consistency

DOI: 10.5281/zenodo.19563630

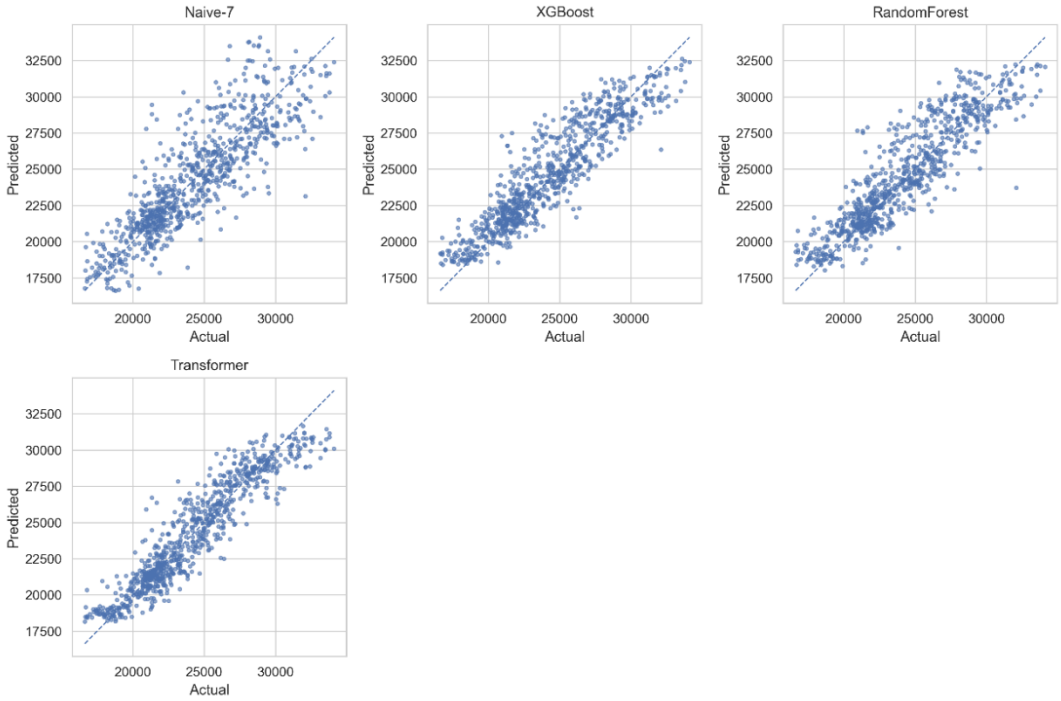


Figure 4. Scatter Comparison

Figure 4 consists of scatter plots comparing the actual demand values with the predicted values generated by the Naive-7, XGBoost, Random Forest, and Transformer models. In each graph, the Y-axis represents the predicted values, while the X-axis represents the actual values. The dotted line indicates the ideal prediction line, where the predicted values are equal to the actual values.

4.4 Residual distribution analysis

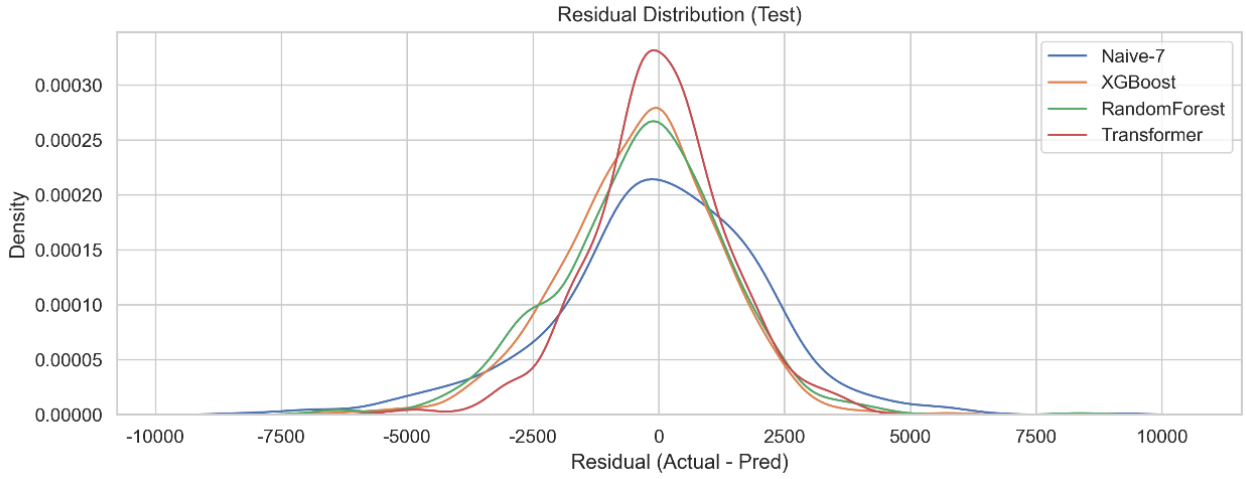


Figure 5. Residual Distribution (Test)

Figure 5 shows the residual distributions of the Naive-7, XGBoost, Random Forest, and Transformer models on the test set. Each curve represents the distribution of residuals (the difference between actual and predicted values) for each model. The X-axis shows the residual values, ranging from negative to positive, while the Y-axis represents the density of the residuals.

The Naive-7 model exhibits a wide and almost symmetrical distribution of residuals, with a peak close to zero. The residual distribution of XGBoost is narrower and more concentrated around zero. The Random Forest residual distribution is also centralized but slightly more dispersed than that of XGBoost. In contrast, the Transformer model shows a sharp peak around zero, indicating smaller residuals compared with the other models.

4.5 Training dynamics and convergence

DOI: 10.5281/zenodo.19563630

Figure 6. Transformer Learning Curve

Figure 6 shows the learning curve of the Transformer model, illustrating the Huber loss over time. The training loss (blue line) decreases rapidly during the initial epochs, while the validation loss (orange line) follows a similar downward trend, though at a slightly slower rate. After approximately five epochs, both losses stabilize, indicating that the model has converged. The consistently small gap between the training and validation losses suggests that the model demonstrates good generalization performance.

4.6 Uncertainty quantification with conformal intervals

Model	PICP	Avg. Width
Naive-7	0.880	6176.18
XGBoost	0.882	4766.23
RandomForest	0.862	5083.41
Transformer	0.974	6585.61

Table 2. Intervals_Stats

Table 2 shows that Transformer model of PICP is highest, 0.974, and Naive - 7 model coverage, lowest 0.880, while the average interval width of the Transformer is the widest, 6585.61.

4.7 Feature interpretability of tree models

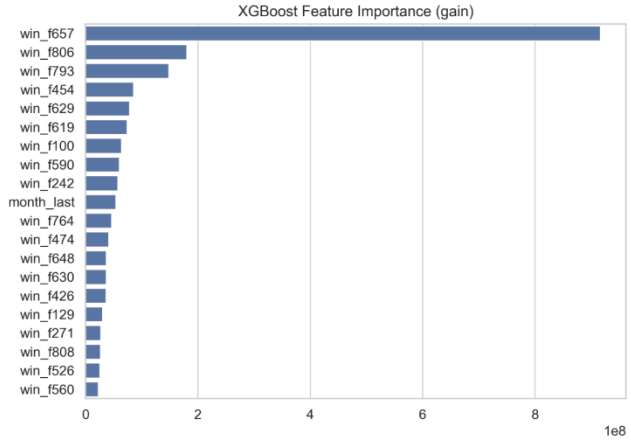


Figure 7. XGBoost Feature Importance (gain)

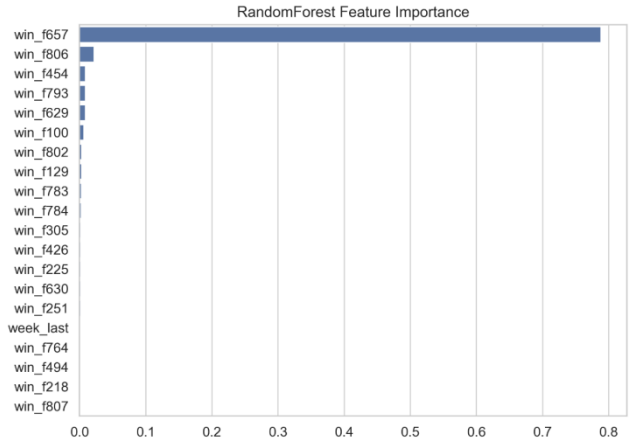


Figure 8. RandomForest Feature Importance

Figure 7 and **Figure 8** show the feature importance of the XGBoost and Random Forest models, respectively.

XGBoost Feature Importance: The bars indicate the relative importance of different features, with *win_f657* being the most significant, followed by *win_f806* and *win_f793*. Other features, such as *win_f454* and *win_f629*, also contribute to the model but are considerably less influential than the top-ranked features.

Random Forest Feature Importance: Similar to XGBoost, *win_f657* is the most important feature, followed by *win_f806*. However, after the top few features, the importance drops more sharply, suggesting that only a small subset of features substantially drives the model's predictions.

4.8 Summary

In summary, the Transformer model outperforms the other models in both point forecasting and uncertainty quantification, achieving the lowest MAE, RMSE, and MAPE values, as well as the highest R-squared values, demonstrating superior accuracy and fit to the data. While XGBoost and Random Forest also perform well, the Transformer model effectively captures long-term dependencies and provides reliable prediction intervals, which are particularly important for power load forecasting. Although there are occasional periods of high fluctuation where some points fall outside the predicted intervals, the Transformer's ability to adapt to sudden variations significantly improves performance compared to the Naive-7 model. Furthermore, the feature importance analysis of the tree-based models indicates that only a few key features

DOI: 10.5281/zenodo.19563630

dominate predictions, suggesting that the Transformer's more comprehensive modeling of multiple features contributes to its robustness across diverse demand scenarios.

5. Conclusions

This study demonstrates the superior performance of a Transformer-based model for power load forecasting, surpassing traditional models such as Naive-7, XGBoost, and Random Forest. The Transformer model not only achieves the lowest MAE, RMSE, and MAPE values in point prediction but also excels in uncertainty quantification using conformal prediction, attaining the highest prediction interval coverage probability (PICP) and providing more reliable forecast intervals.

While tree-based models such as XGBoost and Random Forest also perform well, the Transformer model effectively captures long-term dependencies, adapts to seasonal variations, and generates dynamic prediction intervals, making it a robust solution for increasingly complex and dynamic energy systems.

Furthermore, feature importance analysis indicates that tree-based models rely on a limited set of key features, whereas the Transformer benefits from a broader, more integrated approach to modeling. This study highlights the potential of advanced machine learning models, particularly the Transformer, to improve both the accuracy and reliability of power demand forecasting. The incorporation of quantified uncertainty adds significant value for decision-making, providing actionable insights into future energy demand scenarios—critical for effective grid management, policy planning, and the integration of renewable energy sources.

References

1. Jiang, P., Li, R., Lu, H., Chen, X., & Wang, Y. (2020). Modeling of electricity demand forecast for power system. *Neural Computing and Applications, 32*(10), 6857–6875. https://doi.org/10.1007/s00521-019-04153-5
2. Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Heshmati, A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems, 8*(2), 411–447. https://doi.org/10.1007/s12667-016-0203-y
3. Iranmanesh, H., Abdollahzade, M., & Miranian, A. (2012). Mid-term energy demand forecasting by hybrid neuro-fuzzy models. *Energies, 5*(1), 1–21. https://doi.org/10.3390/en5010001
4. Khairalla, M. A., Ning, X., Al-Jallad, N. T., & El-Faroug, M. O. (2018). Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. *Energies, 11*(6), 1605. https://doi.org/10.3390/en11061605
5. Sheikhan, M., & Mohammadi, N. (2013). Time series prediction using PSO-optimized neural network and hybrid feature selection algorithm for IEEE load data. *Neural Computing and Applications, 23*(5–6), 1185–1194. https://doi.org/10.1007/s00521-012-0980-8
6. Tilahun, F. B., Bhandari, R., & Mamo, M. (2019). Supply optimization based on society's cost of electricity and a calibrated demand model for future renewable energy transition in Niger. *Energy, Sustainability and Society, 9*(31). https://doi.org/10.1186/s13705-019-0217-0
7. van Ruijven, B., van der Sluijs, J. P., van Vuuren, D. P., & de Vries, B. (2010). Uncertainty from model calibration: Applying a new method to transport energy demand modelling. *Environmental Modeling & Assessment, 15*(2), 175–188. https://doi.org/10.1007/s10666-009-9200-z

DOI: 10.5281/zenodo.19563630

DOI: 10.5281/zenodo.19563630