

Article

Short-Term Electrical Load Forecasting at a 15-Minute Resolution: A Benchmarking Study Using a Rolling-Window Training Approach Against Official TSO Forecasts

Kamil Misiurek , Tadeusz Olkusi  and Janusz Zyśk 

Faculty of Energy and Fuels, AGH University of Krakow, 30-059 Kraków, Poland; olkusi@agh.edu.pl (T.O.); jazysk@agh.edu.pl (J.Z.)

* Correspondence: kmisiurek@agh.edu.pl

Abstract

Modern power systems require increasingly precise forecasts of electricity consumption, which are crucial for effective planning, reducing the risk of power shortages, and integrating renewable energy sources. This article presents the results of comparative benchmarking studies on short-term load forecasting (STLF) at a 15 min resolution. The study uses the rolling-window training method, and the results were compared with the official forecasts of the Transmission System Operator (TSO). The analysis is based on time series and machine learning methods, with the aim of improving the operational accuracy necessary for stable and secure management of the power management system.

Keywords: short-term load forecasting; time series analysis; machine learning; electrical load forecasting; Transmission System Operator (TSO)

1. Introduction

Accurate forecasting of electricity consumption has become increasingly critical in modern power systems. It enables more efficient system planning, reduces the risk of capacity shortages, and supports the integration of renewable energy sources [1]. Load forecasts underpin a wide range of operational decisions, including system management, planning, scheduling, and load dispatch. Consequently, high forecasting accuracy is essential to ensure the reliable, secure, and stable operation of power systems. Conversely, inaccurate forecasts may lead to substantial economic losses for power companies by increasing operating costs. Previous studies indicate that even a 1% increase in forecasting error can result in additional annual operating expenses of approximately USD 10 million every year in China [2].

A review of the relevant literature highlights two key aspects. First, due to the inherently stochastic nature of electricity demand, no universal forecasting method exists that performs optimally under all conditions. Second, the ever-increasing demands for higher accuracy have driven the emergence of novel methodological approaches [3].

Forecasting objectives are closely linked to the considered time horizon. Very-short-term and short-term forecasting play a crucial role in effective demand management and in assessing the potential for energy storage utilization. In this context, these terms typically describe a time window below one day, ranging from minutes to several hours ahead. Accurate day-ahead or multi-day forecasts help mitigate peak loads and improve operational efficiency [4]. Prior research shows that short-term and very short-term load



Academic Editor: Jesús Polo

Received: 11 June 2026

Revised: 2 July 2026

Accepted: 3 July 2026

Published: 7 July 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

forecasting is particularly important for demand planning in regional power systems and microgrids serving large individual consumers, such as buildings (e.g., schools, hotels, offices), infrastructure facilities, and industrial enterprises. Medium-term forecasting, in turn, is essential for generation planning and power grid maintenance. In particular, monthly forecasts are critical for the operation of thermal power plants, where they play a central role in fuel management and energy trading [5]. This horizon typically spans from one month to one year. Long-term forecasting is primarily used for strategic planning of power system development at the national and sectoral levels, as well as for investment planning and major infrastructure refurbishment. This horizon typically covers a period from one year to several years, and often up to a decade or more. Accurate long-term forecasts are also fundamental for designing effective environmental strategies, including industrial decarbonization scenarios [6].

A critical turning point occurred on 30 September 2025, when the European day-ahead electricity markets within the Single Day-Ahead Coupling (SDAC) framework transitioned from hourly to 15 min trading intervals. This regulatory shift was introduced to better reflect the physical reality of power systems increasingly dominated by variable renewable energy sources, such as wind and solar power. Under this new regime, data granularity increases fourfold, making short-term load forecasting (STLF) at a 15 min resolution a critical operational necessity [7].

In this context, official forecasts published by Transmission System Operators (TSOs) serve as a market benchmark for system planning and security. However, given the growing penetration of stochastic and renewable generation and evolving consumption patterns, an important question arises as to whether traditional models employed by TSOs remain the most effective solution, or whether advanced data-driven approaches can deliver superior performance [8].

Therefore, this article aims to bridge the gap between theoretical model development and the practical market application of short-term load forecasting models. The novelty of this study lies in the operational benchmarking of 15 min short-term load forecasting models against the official forecast of the Polish Transmission System Operator under a rolling-window evaluation framework. Unlike many previous studies that validate forecasting models mainly against internal statistical baselines, this work uses the TSO forecast as an external industry benchmark and evaluates the extent to which independently developed univariate models can approach operational forecasting accuracy using only publicly available historical load and calendar information.

The main research contributions of this study are fourfold. First, the paper provides a 15 min resolution benchmarking framework aligned with the current European market structure, in which sub-hourly forecasting has become increasingly important. Second, it compares classical, machine learning, and deep learning models under the same rolling-window training and testing conditions. Third, it quantifies the performance gap between open univariate forecasting models and the official TSO forecast using MAE, RMSE, MAPE, and R^2 . Fourth, it evaluates model robustness under non-standard demand conditions, including weekends and public holidays, and verifies the statistical significance of improvements over the Seasonal Naïve benchmark using the Wilcoxon signed-rank test.

2. Literature Review

Very short-term electricity load forecasting is a critical process used to predict electricity demand in intervals as brief as 30 min, usually performed one day ahead.

Reference [9] discusses the methodology for very short-term electric load forecasting using the Interval Type-1 Fuzzy Inference System (IT-1 FIS). The forecasting is conducted one day in advance at 30 min intervals, which is crucial for operational planning, main-

tenance, and cost optimization in energy management. The study focuses on the impact of different time difference settings on forecast accuracy. Traditionally, forecasting is performed by analyzing historical data from the same day of the week in previous years. The authors propose an alternative approach in which, instead of analyzing past Fridays, data from previous weekdays (Thursday, Wednesday, and Tuesday) are used. The results indicate that the classical method (using past Fridays) produced a lower forecast error (0.4695%) compared to the alternative approach (0.6123%).

Ref. [10] investigated residential load forecasting at 30 min intervals using Long Short-Term Memory (LSTM) networks. The study highlights that LSTM outperforms traditional methods (e.g., KNN, BPNN, and ELM) in handling the high volatility of individual household consumption. A key finding is that aggregating individual forecasts significantly improves accuracy, reducing RMSE by 10–15% compared to direct total-load forecasting. The authors conclude that LSTM effectively captures complex temporal dependencies, suggesting that future improvements could stem from hybrid models and the inclusion of weather data.

The article [11] introduces a hybrid CNN–LSTM framework for residential energy forecasting. The model utilizes Convolutional Neural Networks (CNN) for spatial feature extraction and noise reduction, while the LSTM component captures long-term temporal dependencies. Using a 60 min sliding window, the hybrid approach achieved a MAPE of 34.84%, significantly outperforming linear regression and standalone LSTM models. The results demonstrate that integrating spatial and temporal feature extraction enhances the robustness of forecasts across various time resolutions.

Ref. [12] proposes a deep LSTM model for short-term load forecasting in Spain, utilizing a high-resolution 10 min dataset spanning 9.5 years. The study investigates hyperparameter optimization using Random Search and a novel Coronavirus Optimization Algorithm (CVOA). With a 168-step historical look-back window, the LSTM achieved a MAPE of 1.4472% for a 4 h horizon, outperforming advanced architectures like Temporal Fusion Transformers (TFT) and gradient boosting. The analysis also highlights that forecasting accuracy significantly degrades during holiday periods (e.g., Christmas) and seasonal transitions, identifying these as key challenges for model robustness.

The study [13] focuses on enhancing short-term load forecasting by developing hybrid models based on the CatBoost and XGBoost algorithms. The researchers optimized these models using three distinct metaheuristic algorithms: the Arithmetic Optimization Algorithm (AOA), Phasor Particle Swarm Optimization (PPSO), and the Aquila Optimizer (AO). The XGBoost–AOA hybrid model demonstrated the best performance on the testing dataset, offering high accuracy and significantly shorter runtimes than CatBoost counterparts. Sensitivity and importance analyses identified temperature as the most influential variable affecting electricity load patterns. The inclusion of temporal factors, specifically the month variable, also showed a notable impact on forecasting precision.

Accurate load forecasting at sub-hourly time resolutions has gained increasing attention in recent years due to the evolving operational requirements of modern power systems. While traditional load forecasting studies have predominantly focused on hourly resolution, the growing penetration of variable renewable energy sources, the expansion of intraday electricity markets, and the deployment of energy storage systems have significantly increased the need for higher temporal granularity. In particular, 15 min load forecasting has become a critical component of system operation, as it better captures intra-hour variability and reflects the actual dynamics of electricity demand [14].

High-resolution load forecasting poses several challenges that distinguish it from conventional hourly forecasting. Sub-hourly time series are characterized by increased short-term volatility, a higher noise-to-signal ratio, and stronger sensitivity to calendar

effects, such as weekdays, weekends, and public holidays. Additionally, high-resolution datasets are more susceptible to structural inconsistencies caused by Daylight Saving Time (DST) transitions, which require careful preprocessing to maintain temporal continuity. These characteristics make sub-hourly forecasting a non-trivial extension of hourly models and often necessitate dedicated modeling strategies [15].

Despite the growing body of literature, the majority of existing studies evaluate model performance using internal benchmarks or synthetic baselines, with limited emphasis on comparisons against real operational forecasts used by system operators. As a result, the practical relevance of many high-resolution forecasting models remains insufficiently assessed. This gap is particularly evident in the context of 15 min load forecasting aligned with current market operation frameworks, underscoring the need for systematic benchmarking against official Transmission System Operator forecasts.

3. Materials and Methods

This section describes the data sources, preprocessing procedures, forecasting framework, and evaluation methodology employed in this study. The proposed approach is designed to ensure full reproducibility and objective comparison of forecasting models under realistic operational conditions. Particular emphasis is placed on high-resolution (15 min) load data aligned with current market operation structures, as well as on the use of official Transmission System Operator forecasts as an industry-standard benchmark.

3.1. Data Description

The primary dataset used in this benchmarking study comprises actual, ex-post historical electricity load data and official STLF for the Polish Power System, retrieved from the public repository of Polskie Sieci Elektroenergetyczne S.A. (PSE), the national Transmission System Operator (TSO) in Warsaw, Poland. The dataset spans 19 months, from 14 June 2024 to 25 January 2026. Models were trained on data from 14 June 2024 to 30 November 2025 and evaluated on an out-of-sample test period from 1 December 2025 to 25 January 2026. This timeframe provides a robust basis for evaluating the models across various seasonal patterns, including the high-demand winter peaks and the specific load characteristics of the summer periods.

The data resolution is 15 min, which corresponds directly to the Imbalance Settlement Period (ISP) currently implemented in the Polish electricity market. This high granularity is crucial for modern power system operations, as it allows for a more precise capture of intra-hour load fluctuations compared to the traditional hourly resolution. The dataset includes two main time series:

- Actual System Load: The real-time metered electricity consumption in the Polish grid.
- TSO Forecast: The official day-ahead load forecast published by PSE, which serves as the industry benchmark for this study.

3.2. Data Preprocessing

The quality of the retrieved dataset was high, and a preliminary audit confirmed that there were no missing observations (NaN values) within the analyzed period. To ensure the robustness of the forecasting models, the following preprocessing steps were applied:

- Outlier Detection: Although the dataset provided by the TSO was pre-filtered, the series were screened for extreme anomalies using a rolling Z-score method. No significant non-technical outliers were identified that would require removal or interpolation.
- Daylight Saving Time (DST) Transitions: To maintain a consistent time series, special attention was paid to the clock changes in October and March. For the transition from

summer to wintertime (October), the additional hour was handled by maintaining the sequence of 15 min intervals. For the winter-to-summer transition (March), the missing 15 min blocks were generated using linear interpolation to ensure a uniform input vector length for the machine learning models.

- Normalization: For models sensitive to the scale of input data (such as Neural Networks), the load values were normalized to a range of [0, 1] or standardized, while the benchmarking against the TSO forecast was performed on the original scale (MW).

3.3. Feature Engineering

Unlike statistical models, machine learning (LightGBM) and deep learning (LSTM) algorithms require proper preparation of input features to effectively identify patterns in high-resolution time series. In this study, three groups of features were implemented.

Energy load exhibits strong daily and weekly seasonality. First, the following information was extracted from the raw timestamp: hour (0–23), minute (0, 15, 30, 45), day of the week (0–6), and weekend information (binary flag).

Secondly, to make the model understand that 11:45 p.m. and 12:00 a.m. are close to each other, a trigonometric sin/cos transformation was applied to the time variables. The transformation was performed according to the following equations: $x_{sin} = \sin(\frac{2\pi t}{T})$ and $x_{cos} = \cos(\frac{2\pi t}{T})$, where t represents the time variable and T is its periodicity (e.g., 24 h). This prevents decision trees and Neural Networks from misinterpreting time as a linear value. Moreover, historical variables have been introduced to inform the model about the dynamics of the system. The training dataset was augmented with the following lag features:

- Lag 24 h—demand value exactly one day ago (96 intervals back).
- Lag 7 d—demand value from a week ago (96×7 intervals back).

These features provide the models with critical information regarding recent trends and long-term seasonal patterns in the power system.

3.4. Forecasting Models

To ensure a comprehensive and methodologically balanced comparison, four forecasting approaches representing different modeling paradigms were considered in this study. The selected models include a Seasonal Naïve benchmark, a classical statistical forecasting model, a tree-based machine learning algorithm, and a deep learning architecture. This combination enables an objective assessment of forecasting performance across varying levels of model complexity under identical data and evaluation conditions. All data-driven models (used in this article) operate under a continuous daily retraining frequency within the rolling-window execution framework. At the start of each simulated operational day, the models were retrained on the updated historical data partition to incorporate the most recent 15 min system dynamics before generating the multi-step-ahead load forecasts for the subsequent 24 h horizon (96 intervals). All computational simulations and rolling-window training loops were executed within a unified hardware environment utilizing an Intel Core i7-1065G7 processor (16 GB RAM) on a 13.5-inch Microsoft Surface Laptop 3. For each simulated day in the rolling framework, the single-step retraining required an average of less than 15 s for LightGBM and approximately 160 s for the LSTM network. Consequently, the total execution time to complete the continuous daily retraining loop across the entire out-of-sample test period was under 15 min for LightGBM and approximately 2.5 h for the deep learning architecture.

3.4.1. Seasonal Naïve Benchmark

As a baseline reference, a seasonal naïve model with a weekly cycle was employed [16]. Unlike the simple daily persistence model, this benchmark assumes that the load at a given 15 min interval on the forecast day is equal to the load observed at the same interval exactly seven days prior. For a time series $Y = \{y_1, y_2, \dots, y_n\}$, mathematically, the multi-step naïve forecast is defined as

$$\hat{y}_{t+k} = y_{t+k-672} \text{ for } k = 1, 2, \dots, 96 \quad (1)$$

where

$\hat{y}_{t+k}$ represents the predicted value at k steps ahead for the target day,
 $y_{t+k-672}$ is the observation from the same time step on the corresponding day of the previous week (672 lags prior, calculated as 7 days $\times$ 96 intervals/day) [17].

Despite its simplicity, this approach is widely used in short-term load forecasting studies and provides a meaningful lower bound for evaluating the added value of more advanced models [18].

3.4.2. Prophet Model

Prophet was applied as a classical statistical forecasting benchmark. It is an additive time series model that decomposes the load signal into trend and seasonal components and is designed to robustly capture daily and weekly consumption patterns. Mathematically, the Prophet model is formulated as a decomposable time series consisting of three main components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2)$$

where

$g(t)$ represents the trend function (modeling non-periodic changes),
 $s(t)$ represents periodic changes (seasonality), modeled using a Fourier series,
 $h(t)$ represents the effects of holidays or irregular events,
 ϵ_t is the error term assumed to be normally distributed [19].

In this study, to capture institutional electricity demand drops, the holiday effect $h(t)$ was integrated using a dataframe containing key Polish public holidays and seasonal milestones (New Year's Day, Christmas Eve, Christmas Day, Second Day of Christmas, and New Year's Eve). To allow the model to dynamically adjust to these load drops, the holiday prior scale parameter (`holidays_prior_scale`) is set to 15. Furthermore, the model is configured with full Fourier-series-based seasonal components, incorporating active daily, weekly, and yearly seasonalities (`daily_seasonality = True`, `weekly_seasonality = True`, `yearly_seasonality = True`). The trend changepoint prior scale (`changepoint_prior_scale`) is maintained at its robust default value of 0.05 to prevent overfitting while allowing automatic detection of baseline consumption shifts during the rolling-window execution.

In this study, Prophet relies exclusively on historical load observations and does not require explicit feature engineering, making it a suitable and interpretable baseline for short-term forecasting tasks [20].

3.4.3. LightGBM Model

A tree-based gradient boosting regression model implemented using the Light Gradient Boosting Machine (LightGBM) framework was employed as the primary machine learning approach. LightGBM constructs an ensemble of decision trees in a sequential boosting process, enabling the model to capture complex nonlinear relationships between the target variable and engineered input features. The model utilizes calendar-based features,

lagged load values, and rolling statistical indicators derived strictly from past historical data. To capture the recent volatility and baseline demand shifts, two rolling features were computed over a 24 h lookback window (96 steps) using only past observations prior to the forecast execution time: the rolling mean and the rolling standard deviation. The forecasting model is formulated as an ensemble of K decision trees. For a dataset with n samples, the predicted load value $\hat{y}_i$ for a given input vector x_i is obtained by summing the contributions of individual trees:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F} \quad (3)$$

where $\mathcal{F}$ is the space of regression trees. The model is trained in an additive manner. At each iteration t , a new tree f_t is added to minimize the objective function $\mathcal{L}^{(t)}$, which consists of a loss function l measuring the difference between the target y_i and the prediction from the previous step $\widehat{y}_i^{(t-1)}$ and a regularization term Ω .

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (4)$$

To ensure optimal convergence speed and adaptability within the rolling-window infrastructure, the training partition for LightGBM is constrained to a moving historical lookback window of all historical days preceding each forecasted day. Missing lag or statistical observations are dynamically imputed using the historical mean baseline computed from the active training frame. The ensemble architecture is configured with a maximum of 500 decision tree estimators (`n_estimators` = 500) and trained with a contraction learning rate of 0.05. The structural complexity of individual trees is governed by restricting the maximum tree depth to 8 (`max_depth` = 8) and the maximum number of leaves to 31 (`num_leaves` = 31), enforcing a minimum of 20 samples per leaf (`min_child_samples` = 20). To enhance generalization and guard against structural overfitting, stochastic row sub-sampling (`subsample` = 0.8) and column sub-sampling (`colsample_bytree` = 0.8) are enforced at each split. The tree generation process is guided by an objective gain metric and monitored by an early stopping validation protocol with a threshold of 50 consecutive iterations (`stopping_rounds` = 50), evaluating the validation matrix [21,22].

3.4.4. LSTM Model

An LSTM neural network was employed as a representative deep learning approach. LSTM networks are specifically designed to model sequential data and capture temporal dependencies over multiple time scales. In this study, the LSTM model was trained using sequences of historical load observations at a 15 min resolution, enabling it to learn short-term dynamics and daily consumption patterns. The inclusion of the LSTM model allows for an assessment of whether deep learning architectures provide a measurable performance advantage over classical statistical and tree-based machine learning models in a high-resolution operational forecasting context [12,23]. Mathematically, the core of the LSTM network is the memory cell, which maintains a cell state C_t regulated by three gates: the forget gate f_t , the input gate i_t , and the output gate o_t . These gates control the flow of information using sigmoid σ and hyperbolic tangent $\tanh$ activation functions. For an input vector x_t at time step t and the previous hidden state h_{t-1} , the transition equations are defined as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

where W and b denote the weight matrices and bias vectors, respectively, $\odot$ represents the element-wise (Hadamard) product, and $\tilde{C}_t$ is the candidate cell state update [24].

In this study, an LSTM network was developed to capture temporal dependencies in the power load data. To ensure numerical stability, all input features and the target variable were normalized to a $[0, 1]$ range using min–max scaling, with an inverse transformation applied to obtain forecasts in the original scale. The input data were reshaped into a three-dimensional [samples $\times$ timesteps $\times$ features], where the sequence length (lookback window) is set to a single timestep (timestep = 1) aligned with eight engineered features. This single-timestep configuration was implemented as a baseline approach for this stage of the research, as the critical temporal dynamics and calendar periodicities are already explicitly provided to the network via the 8 engineered features (such as the sin/cos transformations and the 24 h and 7-day lags). This setup allowed the study to establish a robust and computationally efficient modeling framework within the rolling-window loop without overcomplicating the network architecture at this phase. We acknowledge that this limits the LSTM's native ability to learn sequences from raw data; therefore, developing more advanced architectures with extended multi-timestep lookback windows is already planned as a key focus of our future work. The network architecture consists of an input layer followed by an LSTM layer with 64 units. To mitigate overfitting, a Dropout layer with a rate of 0.2 was implemented, followed by a fully connected Dense layer with 32 neurons. The final output is generated by a single-neuron Dense layer providing the 15 min load forecast. The model was compiled using the Mean Squared Error (MSE) loss function and trained using the Adam optimizer with a learning rate of 0.001. To maintain accuracy over time, the model operates within a rolling-window framework, where it is retrained over a maximum of 30 epochs with a batch size of 64. To facilitate the early stopping mechanism, a validation split of 15% is reserved during each training iteration to monitor validation loss with a patience of three epochs, incorporating the most recent system dynamics and minimizing the impact of non-stationary trends.

3.5. TSO Forecast Benchmark

In addition to the developed statistical and machine learning models, the official day-ahead load forecast published by PSE, the Polish Transmission System Operator, was utilized as an industrial-reference benchmark. These data represent the TSO's actual operational forecasts published in real time. The forecast is updated continuously every 15 min with a rolling horizon covering the subsequent two days, reflecting active intraday grid adjustments based on the most current system states.

Crucially, the TSO forecast constitutes an asymmetric benchmark. While our proposed independent univariate models operate under strict data constraints, relying solely on historical load sequences, the TSO system has continuous real-time access to high-value exogenous data. This includes proprietary grid infrastructure topology, an active commercial market, and advanced professional meteorological consensus forecasts.

Including the TSO forecast in the evaluation allows for verifying not only the statistical superiority of the proposed methods over simple baselines (like Seasonal Naïve) but also their practical applicability in a real-world market environment. If the proposed LightGBM or LSTM models can achieve accuracy comparable to or better than the TSO forecast, without access to the proprietary grid data available to the operator, it demonstrates significant potential for independent market participants.

Hence, this evaluation must not be interpreted as a strict comparison of algorithmic superiority, nor does it imply that the proposed independent models can directly replace the TSO infrastructure. Instead, the TSO forecast serves strictly as a high-fidelity, real-world operational reference. Achieving competitive accuracy under such strict, univariate data constraints underscores the practical utility of these open models for independent market participants who operate without access to proprietary system topology or real-time meteorological consensus.

3.6. Evaluation Metrics

To quantitatively assess the forecasting performance of the developed models, four standard statistical metrics were employed: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). These metrics are widely adopted in energy forecasting literature [25], enabling a comprehensive comparison of model accuracy and robustness.

MAE measures the average magnitude of errors in the predictions, without considering their direction. It provides a straightforward interpretation of the error in the original unit (MW):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (11)$$

where y_i represents the actual load value, $\hat{y}_i$ is the forecasted value, $\bar{y}$ is the mean of the actual values, and n denotes the total number of samples in the evaluation set (test set).

RMSE represents the standard deviation of the prediction errors. Due to the squaring term, RMSE penalizes large errors more heavily than MAE, making it particularly useful for detecting instances where the model deviates significantly from the actual load:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

MAPE expresses the prediction error as a percentage of the actual value. It is the industry-standard metric for comparing forecasting performance across different grid sizes or time periods:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% \quad (13)$$

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, providing a measure of goodness of fit

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

These metrics are widely adopted in energy forecasting literature [25], enabling a comprehensive comparison of model accuracy and robustness. From the perspective of power system operations, each of these metrics provides a distinct insight into the model's reliability:

- MAE measures the average magnitude of errors in the predictions, without considering their direction. It provides a straightforward interpretation of the error in the original unit (MW). For a grid operator, MAE directly correlates with the physical volume of energy that must be balanced in the real-time market; thus, it is a key indicator of the potential balancing costs.
- RMSE represents the standard deviation of the prediction errors. Due to the squaring term, RMSE penalizes large errors more heavily than MAE. In this study, RMSE is treated as a critical safety metric. High RMSE values indicate the occurrence of large, infrequent forecasting spikes, which pose a significant risk to grid stability and can lead to emergency frequency control actions.
- MAPE expresses the prediction error as a percentage of the actual value. It is the industry-standard metric for comparing forecasting performance across different grid sizes or time periods. It allows for benchmarking the developed univariate models against international standards, where an MAPE below 5% is typically considered a high-quality forecast for STLF.
- R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, providing a measure of goodness of fit. In the context of the rolling-window approach, R^2 indicates whether the model has successfully captured the underlying consumption trends and weekly seasonality without the aid of exogenous weather data.

4. Results and Discussion

The following section presents a comprehensive benchmarking analysis of the implemented forecasting models, namely Seasonal Naïve (SNaive), Prophet, LightGBM, and LSTM. The performance of these models is evaluated against the official Transmission System Operator (TSO) forecasts for the Polish power system. All models were tested using a rolling-window approach on 15 min resolution data covering the period from 1 December 2025 to 25 January 2026. This period was deliberately selected to assess model robustness under both standard operating conditions and pronounced holiday-related anomalies, including the Christmas and New Year periods.

The results in Figure 1 demonstrate that the TSO forecast significantly outperforms all other models across every metric. It achieved the highest precision with an MAPE of 2.13% and an R^2 of 0.976, indicating an exceptionally high correlation between the predicted and observed values. Among the machine learning frameworks, LightGBM emerged as the most effective alternative, yielding a MAPE of 3.88% and an MAE of 775.74 MW. Notably, LightGBM demonstrated superior stability compared to the LSTM network, while both models showed similar R^2 values (0.875 and 0.874, respectively), the LSTM exhibited a higher RMSE of 1237.69 MW vs. 903.82 MW for LightGBM. The Prophet model occupied a middle ground with an MAPE of 6.03%, failing to capture the systemic volatility as effectively as the gradient boosting or deep learning approaches. Finally, the SNaive model, used as a baseline for the study, exhibited the lowest comparative predictive power among the evaluated selections, with a MAPE of 7.64% and an R^2 of 0.542. However, it is important to note that from a time-series perspective, an R^2 value of 0.542 still represents a reasonably strong baseline performance. Due to the high temporal autocorrelation inherent in sub-hourly electricity consumption data, a simple weekly persistence approach is capable of capturing over 54% of the underlying system variance without requiring any parametric modeling or computational training.

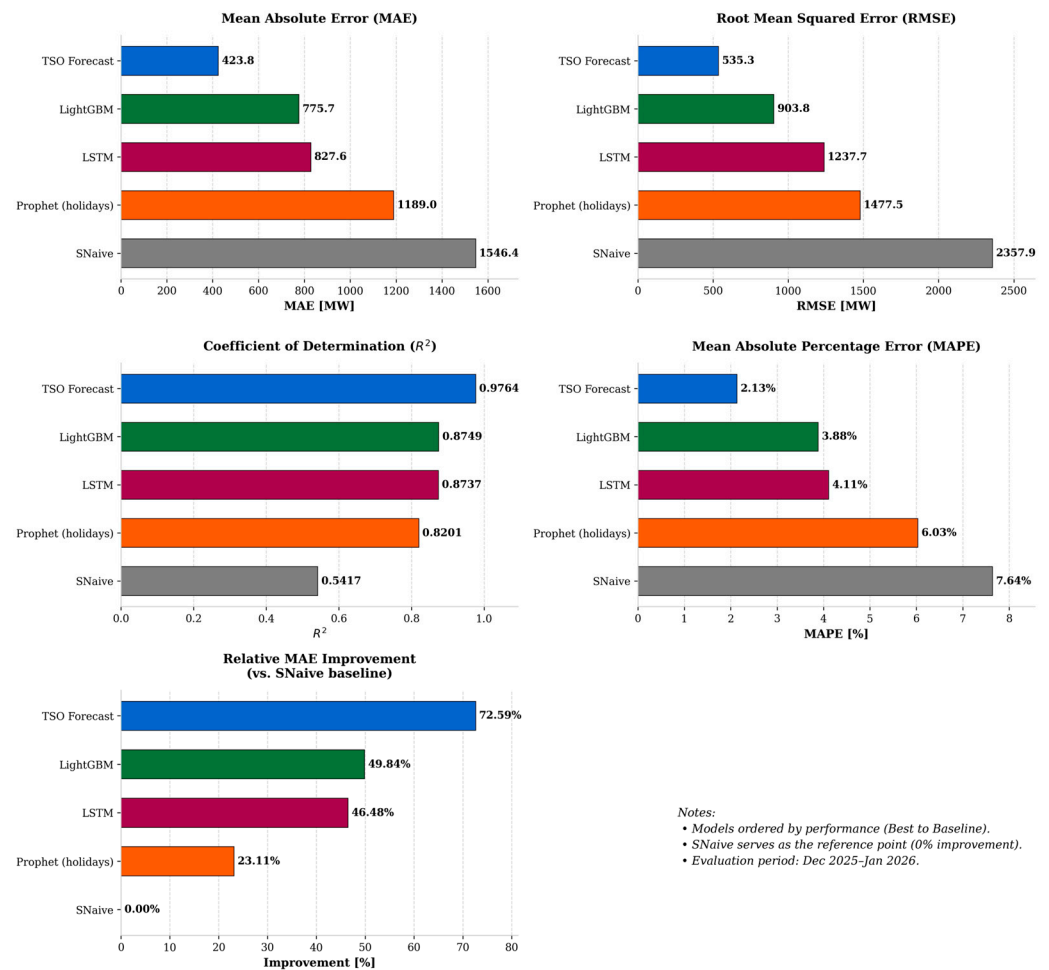


Figure 1. Performance comparison of 15 min load forecasting models (TSO forecast, Seasonal Naïve, Prophet with holiday effects, LightGBM, and LSTM) using MAE, RMSE, R^2 , and MAPE, including MAE improvement relative to the SNaive.

4.1. Statistical Significance Analysis

To assess whether the observed improvements in forecasting accuracy over the baseline model were statistically significant, the Wilcoxon signed-rank test was applied to the absolute forecasting errors. This non-parametric test was selected because the distribution of forecasting errors was not assumed to be normal and because the test period included peak-demand anomalies and holiday-related demand shifts. The analysis was conducted at a 15 min resolution and included 5376 observations covering the period from 1 December 2025 to 25 January 2026. The absolute errors of the LightGBM, LSTM, and Prophet models were compared with those of the SNaive benchmark. In all three comparisons, the resulting p -values were below 0.001, indicating statistically significant differences between the evaluated models and the baseline benchmark. These results support the conclusion that the improvements achieved by LightGBM, LSTM, and Prophet over the Seasonal Naïve benchmark are unlikely to be the result of random variation in the test sample. However, the Wilcoxon signed-rank test was used here as a complementary statistical assessment rather than as the sole basis for ranking the models. Therefore, the relative performance of the models was interpreted jointly with the descriptive accuracy metrics reported in Table 1, including MAE, RMSE, MAPE, and R^2 . Based on these metrics, the TSO forecast remained the most accurate reference across the full test period, while LightGBM achieved the lowest MAE and RMSE among the open forecasting methods. Nevertheless, since a full pairwise statistical comparison between all model combinations was not performed,

the superiority of LightGBM over LSTM and Prophet should be interpreted descriptively, based on the reported error metrics rather than as a separate inferential claim. To better show the practical relevance of the obtained improvements, 95% confidence intervals were calculated for the absolute forecasting errors. The results indicate that the best-performing open model, LightGBM, achieved a narrower error range than LSTM and Prophet. This confirms not only the statistical significance of the results but also a more stable and consistent reduction in forecasting uncertainty. In future work, standardized effect size measures, such as Cohen's d , will be included for multi-year datasets to describe the magnitude of these improvements more precisely.

Table 1. Forecasting accuracy comparison of the evaluated models at a 15 min resolution for the period from 1 December 2025 to 25 January 2026.

Model	MAE [MW]	RMSE [MW]	MAPE [%]	R^2
TSO Forecast	423.78	535.25	2.13%	0.976
LightGBM	775.74	903.82	3.88%	0.875
LSTM	827.60	1237.69	4.11%	0.874
Prophet	1188.96	1477.49	6.03%	0.820
Seasonal Naïve	1546.40	2357.87	7.64%	0.542

4.2. Daily Stability and Error Distribution Analysis

To evaluate the operational reliability of the forecasting models, their daily performance stability and error distributions were examined. Figure 2 presents a comprehensive comparison of the results, including daily MAE, RMSE, R^2 scores, and a comparative boxplot of absolute residuals.



Figure 2. Daily error metrics (MAE, RMSE) and daily R^2 for 15 min load forecasts, with a boxplot of the daily MAE distribution, comparing TSO, SNaive, Prophet, LSTM, and LightGBM.

The temporal analysis of the forecasting models highlights a significant decrease in accuracy during the holiday periods of 24 December, 31 December, and 6 January, etc. While the TSO Forecast demonstrates remarkable resilience with a stable R^2 near 0.976 and a median MAE below 500 MW, competing models exhibit substantial performance degradation due to nonlinear demand shifts. During the Christmas and New Year period, the SNaive model's MAE surged above 8000 MW, with its R^2 collapsing to approximately -11 , while LGBM recorded error peaks nearly reaching 4000 MW. A similar trend occurred around the Epiphany on 6 January, where Prophet and LGBM saw MAE values fluctuate between 2000 MW and 3000 MW as they struggled to model the transition back to standard work-week patterns. Distribution data confirms that although LGBM maintains a competitive median error, it is prone to high-magnitude outliers during these specific holidays, whereas Prophet exhibits a more consistent but higher error baseline around 1000 MW.

4.3. Impact of Holidays and Weekend Volatility

To evaluate model robustness, the testing period was segmented into 32 business days and 24 non-business days (weekends and holidays). The comparative analysis, summarized in Figure 3, reveals distinct behavioral patterns across the forecasting architectures.

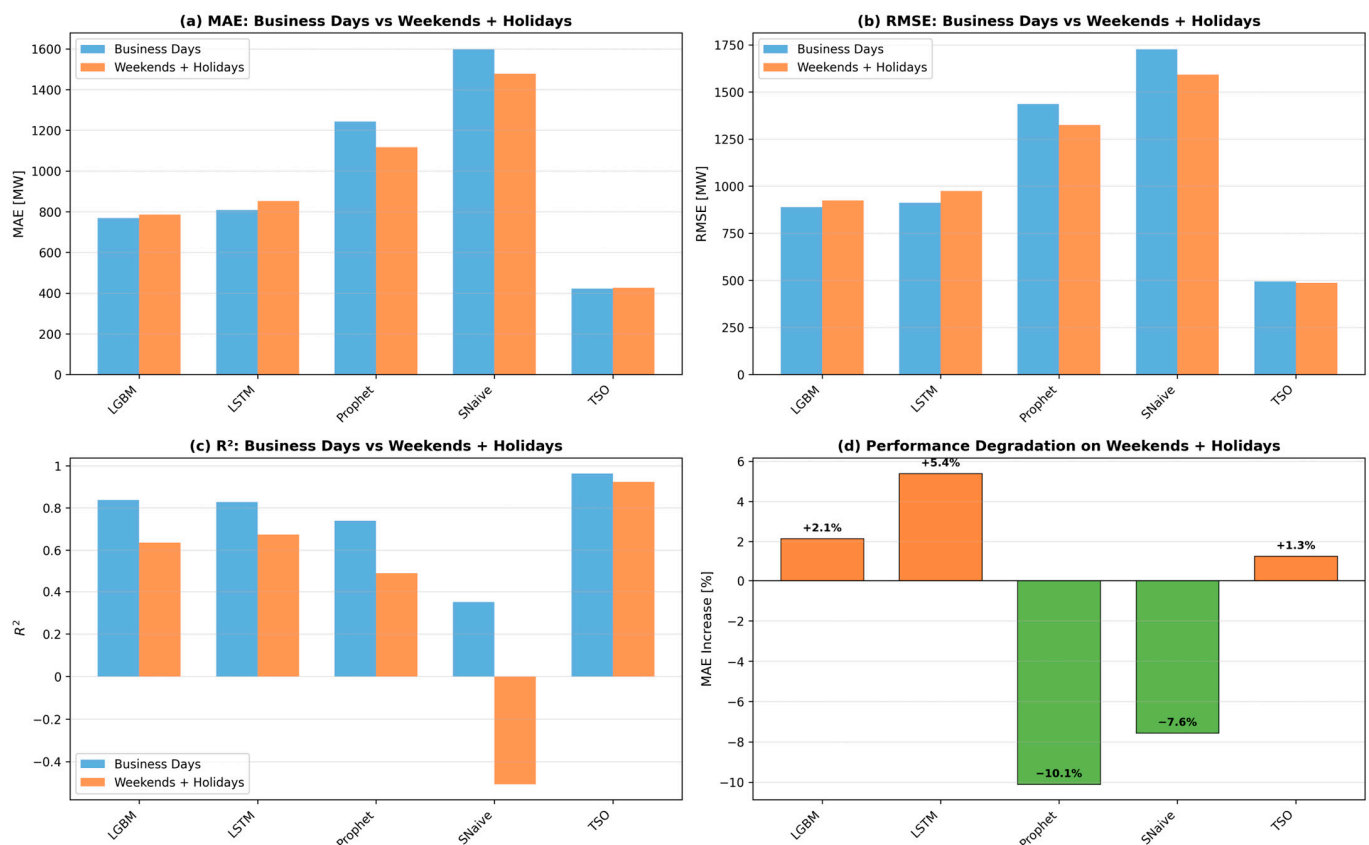


Figure 3. Comparative analysis of forecasting performance between business days and non-business days (weekends and public holidays). The charts present: (a) Mean Absolute Error (MAE) comparison; (b) Root Mean Square Error (RMSE) comparison; (c) R -squared (R^2) stability; (d) Relative MAE change (degradation) across all evaluated models.

The comparative performance analysis, visualized in Figure 3, reveals distinct behavioral patterns across the different modeling architectures when transitioning from standard work-week dynamics to lower-load periods. MAE and RMSE comparisons indicate that while absolute error values for several models decreased on non-business days, this was not necessarily indicative of improved performance, but rather a reflection of the overall

reduction in total system load. A critical observation is the degradation of the R^2 score during these periods, which is most pronounced in the SNaive model, where the coefficient of determination falls to -0.506 , signifying a complete failure to capture holiday-related consumption shifts. The relative performance change, illustrated in the degradation chart (Figure 3d), identifies LSTM as the most sensitive model to non-business day volatility from an empirical standpoint, exhibiting a $+5.4\%$ increase in MAE. While a direct statistical hypothesis test was not performed between the two data-driven models, this descriptive percentage increase indicates that the LSTM architecture experienced a higher relative performance penalty during these specific low-load intervals compared to LightGBM. LightGBM (LGBM) followed a similar trend with a $+2.1\%$ increase in MAE, whereas the TSO Forecast demonstrated exceptional stability, recording only a marginal degradation of $+1.3\%$. Interestingly, Prophet and SNaive models show a statistical decrease in absolute MAE (-10.1% and -7.6% , respectively), yet their significantly lower R^2 values during these times suggest they lack the precision required for reliable holiday forecasting.

Further granular analysis of day-of-week patterns is presented in Figure 4. The MAE heatmap and trend charts highlight that errors for the majority of data-driven models peak mid-week (Wednesday–Thursday) and decrease toward the weekend. Specifically, the LGBM model displays strong performance on Sundays (MAE ~ 410 MW) but struggles significantly on Wednesdays (MAE ~ 1343 MW), likely due to the higher variance in industrial load during the peak business cycle. The weekend degradation analysis by model confirms that when considering purely day-of-week patterns without holiday anomalies, all models, including the TSO benchmark, show a relative reduction in MAE (ranging from -14.1% to -43.1%) compared to standard weekdays.



Figure 4. Day-of-week error patterns for 15 min load forecasting models (TSO, SNaive, Prophet, LightGBM, and LSTM): average MAE for weekdays vs. weekends (based on day-of-week only), weekend–weekday MAE change, day-of-week MAE trend, and MAE heatmap. Weekday holidays are included within the weekday group.

5. Conclusions

This study provided a comprehensive comparative analysis of short-term load forecasting models, focusing on the performance of a Transmission System Operator's (TSO) benchmark against various machine learning and statistical architectures. It is critical to emphasize that the developed independent approaches are intended to serve as robust, open-source forecasting benchmarks for market participants rather than operational alternatives or direct replacements for the institutional TSO forecasting infrastructure, which inherently relies on proprietary grid topology and meteorological consensus. The research highlighted a significant performance degradation across all models during the holiday season, particularly on 24 December, 25 December, 26 December, 31 December, 1 January, and 6 January, where demand patterns deviated sharply from standard business-day profiles. During these volatile periods, the Seasonal Naive model experienced extreme failures, with errors exceeding 8000 MW, whereas the TSO and LightGBM models exhibited greater resilience. Furthermore, the statistical evaluation using the Wilcoxon signed-rank test confirmed that the improvements achieved by the LightGBM, LSTM, and Prophet models over SNaive were statistically significant at the $p < 0.001$ level. Empirically, the TSO forecast achieved the highest accuracy, with a MAPE of 2.13% and an R^2 of 0.976, while LightGBM was the best-performing independently developed model, reaching a MAPE of 3.88% and reducing MAE by approximately 49.8% relative to the Seasonal Naïve benchmark. Among the machine learning models, LightGBM demonstrated superior efficiency and a lower descriptive bias compared to the LSTM network, which empirically showed a higher descriptive sensitivity to non-business-day dynamics with a 5.4% increase in absolute error. These findings suggest that for high-resolution 15 min forecasting, ensemble methods like LightGBM provide a robust balance between computational speed and predictive accuracy. Nevertheless, a key limitation of this study is its strict reliance on univariate historical load sequences and calendar features, leaving the models vulnerable to sudden weather changes. To overcome these constraints, future research should transition toward multivariate forecasting models that integrate exogenous meteorological variables (such as temperature, solar irradiance, and wind speed) across multi-year, cross-seasonal datasets. Exploring more advanced deep learning architectures, including Temporal Fusion Transformers (TFT) and graph neural networks, could provide deeper spatial-temporal modeling capabilities to further enhance forecasting stability during highly anomalous holiday periods. Furthermore, future updates will incorporate probabilistic forecasting capabilities and specific evaluation of ramp-period errors to fully assess the models' readiness for real-time market dispatch safety.

Author Contributions: Conceptualization, K.M., T.O. and J.Z.; methodology, K.M.; software, K.M.; validation, K.M., T.O. and J.Z.; formal analysis, K.M., T.O. and J.Z.; investigation, K.M., T.O. and J.Z.; resources, K.M.; data curation, K.M.; writing—original draft preparation, K.M., T.O. and J.Z.; writing—review and editing, K.M., T.O. and J.Z.; visualization, K.M.; supervision, T.O. and J.Z.; project administration, K.M., T.O. and J.Z.; funding acquisition, K.M., T.O. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out under Subvention from the Faculty of Energy and Fuels, AGH University of Krakow no. 501.00-210000-10000 and IDUB UB-696000-00001-Z01.

Data Availability Statement: The historical electricity load data and official forecasts analyzed during this study are publicly available from the open repository of Polskie Sieci Elektroenergetyczne S.A. (PSE), the Polish Transmission System Operator, and can be accessed at <https://www.pse.pl> accessed on 1 February 2026.

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

References

1. Aydin, B. The Impact of Electricity Consumption on Economic Growth in the European Union. *İZmir İktisat Derg.* **2021**, *36*, 13–26. [CrossRef]
2. Singh, P.; Dwivedi, P.; Kant, V. A hybrid method based on neural network and improved environmental adaptation method using Controlled Gaussian Mutation with real parameter for short-term load forecasting. *Energy* **2019**, *174*, 460–477. [CrossRef]
3. Misiurek, K.; Olkusi, T.; Zyśk, J. Review of Methods and Models for Forecasting Electricity Consumption. *Energies* **2025**, *18*, 4032. [CrossRef]
4. Krstev, S.; Forcan, J.; Krneta, D. An Overview of Forecasting Methods for Monthly Electricity Consumption. *Teh. Vjesn.* **2023**, *30*, 993–1001. [CrossRef]
5. Rani, N.; Aggarwal, S.K.; Kumar, S. Short-Term Load Forecasting Using Combination of Linear and Non-Linear Models. *IEEE Access* **2024**, *12*, 58993–59006. [CrossRef]
6. Adhiswara, R.; Abdullah, A.G.; Mulyadi, Y. Long-term electrical consumption forecasting using Artificial Neural Network (ANN). *J. Phys. Conf. Ser.* **2019**, *1402*, 033081. [CrossRef]
7. Directorate-General for Energy. EU Electricity Trading in the Day-Ahead Markets Becomes More Dynamic. Available online: https://energy.ec.europa.eu/news/eu-electricity-trading-day-ahead-markets-becomes-more-dynamic-2025-10-01_en (accessed on 1 February 2026).
8. Bunn, D.W. Forecasting loads and prices in competitive power markets. In *Proceedings of the IEEE*; IEEE: New York, NY, USA, 2000; Volume 88, pp. 163–169. [CrossRef]
9. Jamaaluddin; Robandi, I.; Anshory, I. Time Difference Calculation Settings for Very Short Term Electric Load Forecasting Using Interval Type-1 Fuzzy Inference System (IT-1 FIS). In *Proceedings of the IOP Conference Series: Materials Science and Engineering, Bengkulu, Indonesia, 17–18 October 2019*; IOP Publishing: Bristol, UK, 2020; Volume 874, p. 012011. [CrossRef]
10. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. In *IEEE Transactions on Smart Grid*; IEEE: New York, NY, USA, 2019; Volume 10, pp. 841–851. [CrossRef]
11. Kim, T.-Y.; Cho, S.-B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
12. Torres, J.F.; Martínez-Álvarez, F.; Troncoso, A. A deep LSTM network for the Spanish electricity consumption forecasting. *Neural Comput. Appl.* **2022**, *34*, 10533–10545. [CrossRef] [PubMed]
13. Zhang, L.; Jánošík, D. Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert Syst. Appl.* **2024**, *241*, 122686. [CrossRef]
14. Dordonnat, V.; Koopman, S.J.; Ooms, M.; Dessertaine, A.; Collet, J. An hourly periodic state space model for modelling French national electricity load. *Int. J. Forecast.* **2008**, *24*, 566–587. [CrossRef]
15. Feng, C.; Shao, L.; Wang, J.; Zhang, Y.; Wen, F. Short-term Load Forecasting of Distribution Transformer Supply Zones Based on Federated Model-Agnostic Meta Learning. In *IEEE Transactions on Power Systems*; IEEE: New York, NY, USA, 2025; Volume 40, pp. 31–45. [CrossRef]
16. Wood, M.; Ogliari, E.; Nespoli, A.; Simpkins, T.; Leva, S. Day Ahead Electric Load Forecast: A Comprehensive LSTM-EMD Methodology and Several Diverse Case Studies. *Forecasting* **2023**, *5*, 297–314. [CrossRef]
17. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, VIC, Australia, 2018.
18. Weron, R. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*; W Wiley Finance Series; John Wiley & Sons: Chichester, UK; Hoboken, NJ, USA, 2006.
19. Taylor, S.J.; Letham, B. Forecasting at Scale. *PeerJ* **2017**. [CrossRef]
20. Amin, N.; Kim, Y.-W.; Kang, C.; Byun, Y.-C. Electricity load forecasting using hybrid prophet-TFT model incorporating weather data. *Energy Rep.* **2026**, *15*, 108846. [CrossRef]
21. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the Advances in Neural Information Processing Systems 30, Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017*; NeurIPS Proceedings: San Diego CA, USA, 2017; pp. 3146–3154. Available online: <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree> (accessed on 1 February 2026).
22. Haben, S.; Ward, J.; Vukadinovic Greetham, D.; Singleton, C.; Grindrod, P. A new error measure for forecasts of household-level, high resolution electrical energy consumption. *Int. J. Forecast.* **2014**, *30*, 246–256. [CrossRef]
23. Qiu, X.; Zhang, L.; Ren, Y.; Suganthan, P.; Amaratunga, G. Ensemble deep learning for regression and time series forecasting. In *Proceedings of the 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL), Orlando, FL, USA, 9–12 December 2014*; IEEE: New York, NY, USA, 2014; pp. 1–6. [CrossRef]

24. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
25. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [[CrossRef](#)]

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.