






Enhancing LSTM-Based Models for Electricity Consumption Forecasting with Bayesian Hyperparameter Tuning

Winita Sulandari^{1*}, Mazka Buana Hidayat¹, Yudho Yudhanto², Dedy Dwi Prastyo³, Salma Dewi Nataya⁴,
Paulo Canas Rodrigues⁵, Muhammad Hisyam Lee⁶

¹ Department of Statistics, Universitas Sebelas Maret, Surakarta 57126, Indonesia

² Department of Informatics Engineering, Vocational School, Universitas Sebelas Maret, Surakarta 57129, Indonesia

³ Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia

⁴ Study Program of Statistics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

⁵ Department of Statistics, Federal University of Bahia, Salvador 40170-110, Brazil

⁶ Department of Mathematical Sciences, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia

Corresponding Author Email: winita@mipa.uns.ac.id

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.121101>

ABSTRACT

Received: 6 September 2025

Revised: 5 October 2025

Accepted: 11 October 2025

Available online: 30 November 2025

Keywords:

electricity consumption forecasting, time series cross-validation, LSTM, BiLSTM, Bayesian optimization, energy management

This study investigates the role of Bayesian hyperparameter optimization in enhancing the forecasting accuracy of recurrent deep learning models, namely Long Short-Term Memory (LSTM) and bidirectional LSTM (BiLSTM), for electricity consumption prediction. The objective is to evaluate the impact of optimization on model performance and generalization across different national contexts. LSTM captures long-range temporal dependencies, while BiLSTM extends this capability by processing sequences in both forward and backward directions. To ensure a robust evaluation, rolling-window time series cross-validation with 32 folds for Germany and 53 folds for Brazil was applied. Bayesian optimization was employed to fine-tune key hyperparameters, enabling the systematic exploration of model configurations. Results show that BiLSTM consistently outperforms LSTM, compared with the non-optimized BiLSTM baseline. BiLSTM Bayesian optimization (BiLSTM-BO) reduces the test Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) by 53.80% and 54.71% for Germany and by 51.26% and 50.56% for Brazil. These findings highlight the advantages of bidirectional processing and data-driven optimization, providing practical guidance for developing accurate and transferable forecasting systems. Although limited to univariate electricity consumption data, the cross-country, dual-frequency evaluation provides new evidence regarding the robustness and adaptability of Bayesian-optimized LSTM-based models for both short-term operations and long-term planning.

1. INTRODUCTION

Electricity consumption forecasting plays a critical role in supporting energy planning, infrastructure development, and grid stability. Accurate demand predictions enable efficient energy management, particularly in the face of increasing variability due to climate change, population growth, and emerging consumption patterns. Many researchers have explored a wide range of approaches for electricity consumption forecasting, from conventional statistical models [1-4] to more recent machine learning and deep learning techniques [5-9]. As electricity systems grow increasingly dynamic, classical models such as Autoregressive Integrated Moving Average (ARIMA) [10] and exponential smoothing [11] often face difficulties in representing the nonlinear and time-dependent behavior of power demand [12].

In recent years, deep learning—especially the Long Short-Term Memory (LSTM) network—has been widely adopted to improve forecasting performance because of its strength in

learning long-term temporal patterns. Studies in the energy field have reported that LSTM models can provide accurate forecasts of electricity usage [13]. A further development, the Bidirectional LSTM (BiLSTM), processes data in both forward and backward directions, offering richer temporal context and demonstrating promising results in related forecasting applications [14-16].

LSTM networks have become one of the popular tools for modern time series forecasting due to their ability to capture long-term dependencies and mitigate vanishing gradient issues, as highlighted by Van Houdt et al. [17]. In the energy domain, Alizadegan et al. [16] showed that BiLSTM outperformed LSTM and traditional models such as ARIMA and Seasonal Autoregressive Integrated Moving Average (SARIMA), achieving lower Root Mean Square Error (RMSE) in short-term forecasts. However, their analysis was limited to a single temporal scale and did not consider hyperparameter optimization

Hyperparameter selection plays a crucial role in achieving

optimal performance. Yet, many studies still rely on fixed or heuristically chosen hyperparameters, which can limit the model’s generalization ability across different time periods or geographical contexts. Common optimization strategies, such as grid search or random search, are widely used but are computationally expensive and time-consuming [18-20]. More recently, Bayesian optimization has emerged as an efficient alternative, offering faster convergence to optimal solutions [21, 22], and improved forecasting performance [23, 24]. Unlike other optimization algorithms, Bayesian optimization (BO) uses a surrogate probabilistic model and an acquisition function [22] to systematically explore the hyperparameter space, making it well-suited for complex models such as LSTM and BiLSTM.

Recent studies have highlighted the benefits of BO-enhanced recurrent models. Michael et al. [23] showed that incorporating BO into stacked LSTM and BiLSTM models improved solar irradiance forecasting under varying weather conditions. Herrera-Casanova et al. [24] applied BO to BiLSTM for photovoltaic power prediction, reporting notable gains over multilayer perceptron and random forest baselines. Marweni et al. [25] further emphasized the value of BiLSTM-BO in handling uncertainty through interval-valued solar forecasting. Similarly, Habtemariam et al. [22] optimized LSTM for wind power prediction, Wang et al. [26] integrated BO with BiLSTM and Kalman filtering for battery state-of-charge estimation, and Li et al. [27] demonstrated improved ECG classification accuracy with BO-enhanced BiLSTM. Collectively, these works confirm the potential of BO in improving model performance across diverse domains. While these studies underline the value of BO for improving recurrent networks, they have primarily focused on domain-specific applications. Evidence on its effect in electricity load forecasting, particularly when comparing LSTM and BiLSTM models across different temporal resolutions, remains limited.

To address this gap, the present study investigates the effect of Bayesian hyperparameter optimization on LSTM and BiLSTM models for electricity load forecasting. Using rolling-window cross-validation, we evaluate daily load data from Germany and monthly consumption data from Brazil, thereby assessing performance across different temporal resolutions and seasonal characteristics. This controlled evaluation isolates the contribution of BO by directly comparing optimized and non-optimized models and quantifying the improvements achieved. The dual-level comparison further provides insights into how temporal granularity and national consumption patterns influence model behavior.

The findings contribute new evidence on the reliability and generalizability of BO-enhanced recurrent models for electricity demand forecasting. This has implications for

practical tuning strategies and the design of adaptable forecasting systems.

The rest of this paper is organized as follows: Section 2 presents the materials and methods, including data and preprocessing, model overview, and forecasting workflow and evaluation. Section 3 presents the results and discusses the findings. Finally, Section 4 provides the conclusion.

2. MATERIALS AND METHODS

This section describes the materials and methodological framework used in this study. It outlines the characteristics of the datasets, preprocessing techniques, model architectures, hyperparameter optimization strategy, and validation procedures applied to evaluate the forecasting performance of LSTM and BiLSTM models enhanced through Bayesian optimization.

2.1 Dataset and preprocessing

This study focuses on time series forecasting of electricity consumption. The datasets were obtained from Kaggle [28], originally provided by ENTSO-E [29] for Germany and from the website of the Energy Research Company (Empresa de Pesquisa Energética, EPE; www.epe.gov.br) for Brazil. A related study that uses and describes the same official data source, though for a different period, is provided in the study [30]. These two country-specific datasets, which differ in frequency and regional characteristics, were selected to demonstrate the broader applicability of the proposed approach in the energy forecasting domain. Descriptions of both datasets are provided in Table 1.

Table 1. Datasets used and corresponding time periods

| Country | Frequency | Period |
|---------|-----------|-------------------------|
| Germany | daily | 01/01/2015 – 31/12/2017 |
| Brazil | monthly | 01/2004 – 04/2025 |

For both datasets, the following standardized preprocessing steps were applied to prepare the data for the neural network models as presented below.

2.1.1 Determining the training and the testing data

To evaluate model generalizability, a rolling-window time series cross-validation approach was employed. In this method, each fold trains the model on a fixed-length window of the time series and tests it on the immediately following period. This preserves the temporal order of the data and prevents information leakage. The training and testing splits are illustrated in Table 2.

Table 2. Training-testing splits

| Country | Fold | Train | Test |
|---------|------|-----------------------|-----------------------|
| Germany | 1 | 01/01/2015–31/10/2017 | 01/11/2017–30/11/2017 |
| | 2 | 02/01/2015–01/11/2017 | 02/11/2017–01/12/2017 |
| | ⋮ | ⋮ | ⋮ |
| | 31 | 31/01/2015–30/11/2017 | 01/12/2017–29/12/2017 |
| | 32 | 01/02/2015–01/12/2017 | 02/12/2017–31/12/2017 |
| Brazil | 1 | 01/2004–12/2019 | 01/2020–12/2020 |
| | 2 | 02/2004–01/2020 | 02/2020–01/2021 |
| | ⋮ | ⋮ | ⋮ |
| | 52 | 04/2018–03/2024 | 04/2024–03/2025 |
| | 53 | 05/2018–04/2024 | 05/2024–04/2025 |

For the daily German dataset, a 32-fold cross-validation was conducted. In each fold, the models were trained on a rolling window of 1,035 days and tested on the subsequent 30 days. This fine-grained validation strategy simulates a realistic operational scenario in which a utility company might recalibrate and deploy its forecasting models every month.

For the monthly Brazilian dataset, a more extensive 53-fold cross-validation was used. Each fold trained the models on a rolling window of 16 years (192 months) and tested them on the following 12 months. This comprehensive validation design ensures that the models are evaluated across a wide range of long-term economic and climatic conditions, offering a robust assessment of their performance.

2.1.2 Data normalization

Two normalization techniques, min-max normalization and z-score standardization, were considered to ensure numerical stability and promote faster convergence during training. Min-max normalization linearly transforms the data to the [0, 1] according to Eq. (1), while the z-score approach rescales the data based on mean and standard deviation of the training set as shown in Eq. (2).

$$X'_t = \frac{X_t - X_{min}}{X_{max} - X_{min}} \quad (1)$$

$$X'_t = \frac{X_t - \mu}{\sigma} \quad (2)$$

where, X'_t is the scaled data of X_t , X_{min} and X_{max} are the minimum and maximum values in the corresponding training dataset, respectively. Notations μ and σ represent the mean and standard deviation of the training data, respectively. The same scaling parameters derived from the training set were applied to the test set to avoid data leakage.

2.1.3 Determining the input-output pairs

Recurrent neural networks require input data to be structured into sequences of a fixed length. Using a sliding window approach, the time-series data was transformed into input/output pairs.

For the daily German data, a sequence length of 7 days was selected to explicitly capture the strong weekly patterns inherent in electricity usage driven by human social and economic activity (e.g., weekday vs. weekend consumption). The model was trained using 7 consecutive days of data to predict consumption on the 8th day.

For the Brazilian monthly dataset, the input sequence was set to 12 months to reflect the clear yearly seasonal pattern influenced by climatic and hydrological factors. Using these 12 consecutive months, the model then predicted electricity consumption for the following (13th) month.

2.2 Models overview

2.2.1 LSTM-based models

The LSTM network is a variant of the Recurrent Neural Network (RNN) developed to overcome the vanishing gradient issue that limits standard RNNs [17, 31]. LSTM incorporates memory cells together with three gating mechanisms: the input gate, the forget gate, and the output gate. These gates regulate the flow of information, allowing the network to retain relevant signals and discard those that are less important as the sequence progresses. With this

mechanism, LSTM networks are able to capture long-term relationships within time-series data, something conventional RNNs often struggle to achieve [32]. The structure of the LSTM unit is shown in Figure 1.

The forget gate decides which parts of the previous cell state should be removed. It uses a sigmoid activation function applied to the combined previous hidden state and current input, generating values between 0 and 1. Values near 0 signal that the information will be discarded, while those near 1 indicate it will be kept. This process is described mathematically in Eq. (3).

$$f_t = [1 + \exp(-W_f[h_{t-1}, X_t] - b_f)]^{-1} \quad (3)$$

where, W_f represents the weight matrix of the forget gate, which controls how the previous hidden state h_{t-1} and the current input X_t contribute to the forget decision. The term b_f denotes the corresponding bias vector, which adjusts the activation threshold and provides additional flexibility for the model to learn when to retain or discard information [17].

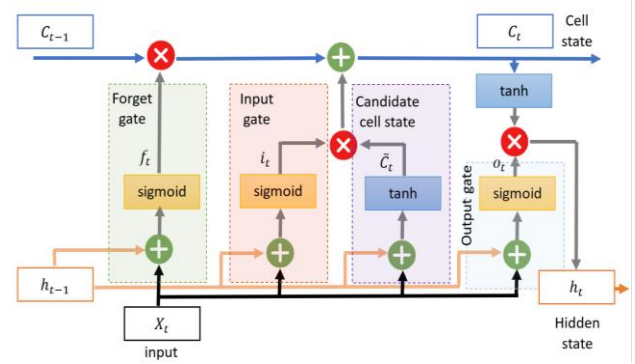


Figure 1. The structure of the LSTM unit

The input gate controls how much new information is incorporated into the cell state. It consists of a sigmoid function that determines the degree of update and a hyperbolic tangent function that produces the candidate values to be added. Together, they update the cell state as defined in Eq. (4) and Eq. (5).

$$i_t = [1 + \exp(-W_i[h_{t-1}, X_t] - b_i)]^{-1} \quad (4)$$

and

$$\tilde{C}_t = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (5)$$

where, $z = W_c[h_{t-1}, X_t] + b_c$. W_c represents the weight matrix associated with the candidate cell state, which determines how much influence the previous hidden state h_{t-1} and the current input X_t have on the candidate values. The term b_c denotes the bias vector, which allows the model to shift the activation function and provides additional flexibility in learning [33].

The cell state is then updated by combining the filtered previous state with the new candidate information. This mechanism enables the LSTM unit to maintain long-term dependencies while gradually incorporating new and relevant inputs, as expressed in Eq. (6).

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (6)$$

Finally, the output gate defines the hidden state at the current time step. It applies a sigmoid function to decide which parts of the updated cell state will contribute to the output, while the cell state is transformed using a hyperbolic tangent function. The resulting hidden and output states are given in Eq. (7) and Eq. (8).

$$o_t = [1 + \exp(-W_o[h_{t-1}, X_t] - b_o)]^{-1} \quad (7)$$

and

$$h_t = o_t \frac{e^{C_t} - e^{-C_t}}{e^{C_t} + e^{-C_t}} \quad (8)$$

where, W_o and b_o are the weight matrix and bias vector of the output gate, respectively, and C_t is the cell state obtained from Eq. (6).

Furthermore, the concept of bidirectional recurrent neural networks was first proposed by Schuster and Paliwal [34]. This idea was later extended by Graves and Schmidhuber [35], who introduced BiLSTM by integrating LSTM units into the bidirectional architecture, where the input sequences are processed in both forward and backward directions. As shown in Figure 2, a BiLSTM consists of two parallel LSTM layers: one processes the input sequence in the forward direction (past to future), while the other processes it in the backward direction (future to past). The outputs from both directions are then combined, enabling the network to incorporate information from the entire sequence at each time step. This bidirectional structure enhances the model's ability to capture complex temporal relationships and dependencies that may not be fully represented by unidirectional models [36].

2.2.2 Bayesian hyperparameter optimization

BO was employed to systematically tune key hyperparameters. The optimization process with the objective of minimizing mean square error as the validation loss. The search space included the number of units in the recurrent layer, learning rate, batch size, and dropout rate. Each trial was evaluated using cross-validation to ensure the robustness of selected parameters. In this work, we considered a model architecture with the hyperparameters described in Table 3.

BO was implemented using `gp_minimize` (scikit-optimize) with a Gaussian process surrogate and expected improvement acquisition function. The search space included units (32–256), learning rate (10^{-4} – 10^{-2} , log-uniform), dropout (0.10–0.50), and batch size (16–128). Each fold used 10 evaluations, with validation loss as the objective; early stopping (patience = 20) ensured convergence. The best configuration was then retrained with early stopping (patience = 50).

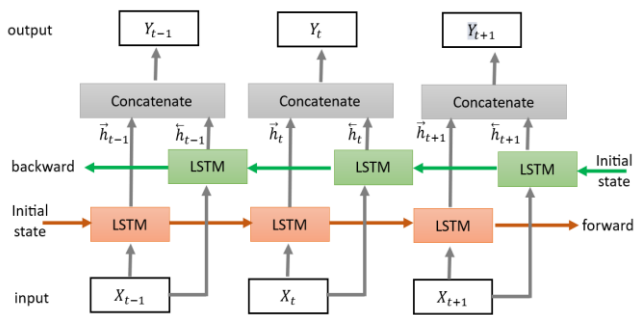


Figure 2. The structure of BiLSTM unit

Table 3. LSTM-based model architecture

| Parameter | Non-BO | With BO |
|-----------------|--|--|
| Recurrent units | 50 | Range: 32 to 256 |
| Layers | 2 LSTM layers 2 Dropout layers 1 Dense layer | 2 LSTM layers 2 Dropout layers 1 Dense layer |
| Dropout rate | 0.0 | Range: 0.1 to 0.5 Logarithmic |
| Learning rate | 0.001 | Range: 1×10^{-4} to 1×10^{-2} |
| Batch size | 32 | Range: 16 to 128 |
| Optimizer | Adam | Adam |
| Callback | Early Stopping (monitors val_loss) 20 epochs | Early Stopping (monitors val_loss) 20 epochs |
| Epoch | 100 | 200 |

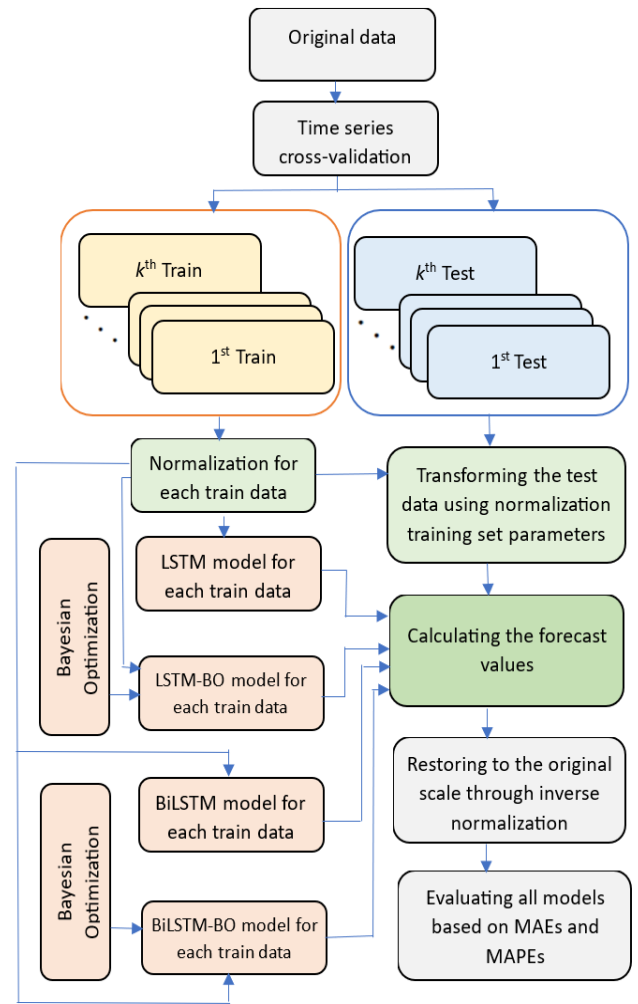


Figure 3. Experimental workflow used in this study

2.3 Experimental workflow and evaluation

The workflow of this work is represented in Figure 3. Performance of each model was evaluated using the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), calculated for each fold and then averaged across all folds for each country.

3. RESULT AND DISCUSSION

The characteristics of the datasets used in this study reveal

notable differences in temporal frequency and consumption patterns, which can influence how the forecasting models perform. The daily frequency for Germany and the monthly frequency for Brazil are illustrated in Figures 4 and 5. These differences should be taken into account when comparing forecasting performance across the two datasets.

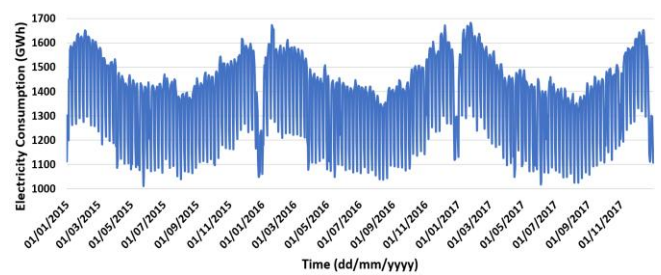


Figure 4. Daily electricity consumption in Germany for the period January 1, 2015 to December 31, 2017

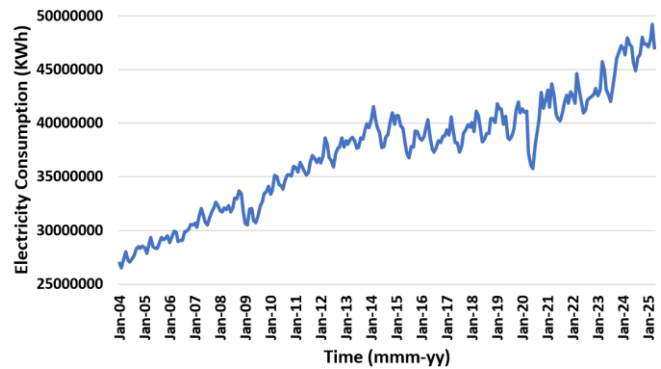


Figure 5. Monthly electricity consumption in Brazil for the period January 2004 to April 2025

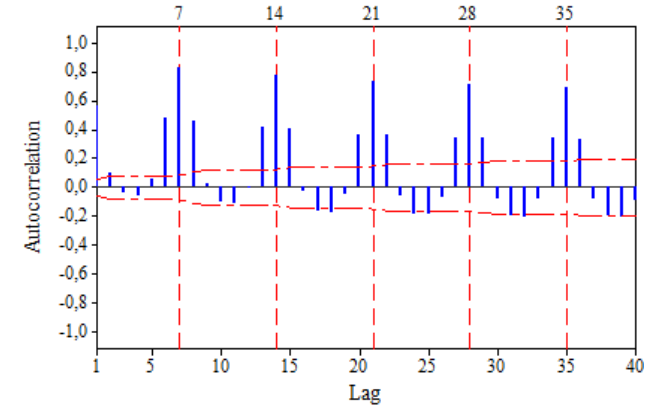


Figure 6. The ACF of daily German electricity consumption

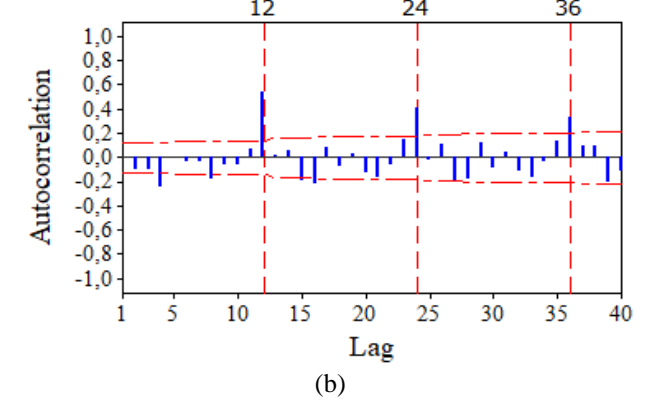
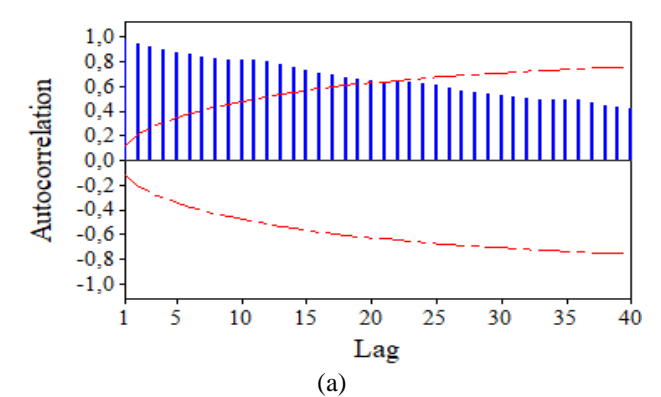


Figure 7. The ACF of the monthly Brazil electricity consumption (a) the original series; (b) the first difference series

Figure 4 shows a clear repetitive pattern, while Figure 5 has a linear upward trend with no obvious repeating pattern. To better understand their temporal structure, the seasonal periods of each dataset were examined using the autocorrelation function (ACF) analysis, which highlights repeating patterns and periodicity (see Figures 6 and 7).

Figure 6 shows a repeating pattern every 7 lags, indicating that the German dataset exhibits weekly seasonality with a period of 7 days. In contrast, Figure 7(a) shows an ACF that gradually decreases, supporting the conclusion that the Brazilian monthly dataset follows a trend pattern. By removing this trend through differencing, the seasonal pattern becomes apparent, with a seasonal period of 12 months, as shown in Figure 7(b).

Based on the identified seasonal periods, the input sequence length for the LSTM and BiLSTM models was set to 7 for the German dataset and 12 for the Brazilian dataset. These values correspond to one full seasonal cycle, allowing each model to capture the dominant periodic behavior effectively. The choice of input length was guided by the autocorrelation function (ACF) of each dataset, which revealed distinct seasonal patterns: a strong weekly cycle for the German daily series and a pronounced annual cycle (12-month lag) for the Brazilian monthly data. Following common practice in short-term load forecasting (e.g., studies [37, 38]), one complete seasonal cycle was used to represent the most influential temporal dependencies while maintaining computational efficiency. Using longer input windows could introduce redundant information and increase model complexity without substantial gains in accuracy. Moreover, in this study, the main objective was to compare model performance with and without BO. Therefore, keeping the input sequence length consistent across experiments ensured that any observed differences in forecasting accuracy could be attributed primarily to the optimization process rather than to variations in model input design. Aligning the input sequence length with the dominant seasonal period enables recurrent architectures to better learn cyclical dependencies, thereby improving forecasting accuracy for time series with strong periodic components.

3.1 Result

In this work, both min–max normalization and z-score standardization were evaluated, as recommended in the literature. Tables 4-7 present the average MAE and MAPE

values for the training and testing data of all models for Germany and Brazil, based on both min–max normalization and z-score standardization. These values were computed across all folds using the rolling-window time series cross-validation described in Section 2. The comparative results (Tables 4–7) show that neither method consistently outperformed the other across datasets, models, or metrics: standardization tended to perform better for the Germany dataset, while the Brazil dataset produced mixed outcomes. For the sake of consistency across experiments, min–max normalization was ultimately adopted in this study, but we note that the choice of normalization method appears to be dataset-dependent and may warrant further investigation in future work.

Table 4. Comparison of MAEs (min-max normalization)

| Model | Germany | | Brazil | |
|-----------|---------|--------|---------|-----------|
| | Train | Test | Train | Test |
| LSTM | 42.55 | 104.42 | 645,551 | 2,402,183 |
| LSTM-BO | 48.71 | 92.81 | 466,363 | 1,368,288 |
| BiLSTM | 34.91 | 81.37 | 621,617 | 2,383,356 |
| BiLSTM-BO | 13.46 | 37.59 | 376,315 | 1,161,574 |

Table 5. Comparison of MAPEs (min-max normalization)

| Model | Germany | | Brazil | |
|-----------|---------|------|--------|------|
| | Train | Test | Train | Test |
| LSTM | 3.62 | 6.55 | 1.75 | 5.80 |
| LSTM-BO | 3.18 | 7.31 | 1.24 | 3.13 |
| BiLSTM | 2.66 | 5.60 | 1.65 | 5.40 |
| BiLSTM-BO | 1.00 | 2.70 | 0.99 | 2.67 |

Table 6. Comparison of MAEs (standardized data)

| Model | Germany | | Brazil | |
|-----------|---------|-------|---------|-----------|
| | Train | Test | Train | Test |
| LSTM | 11.06 | 43.48 | 658,285 | 2,507,053 |
| LSTM-BO | 9.85 | 51.44 | 452,682 | 1,807,252 |
| BiLSTM | 8.01 | 42.41 | 753,338 | 1,870,825 |
| BiLSTM-BO | 7.61 | 33.95 | 382,932 | 1,440,265 |

Table 7. Comparison of MAPEs (standardized data)

| Model | Germany | | Brazil | |
|-----------|---------|------|--------|------|
| | Train | Test | Train | Test |
| LSTM | 0.81 | 2.94 | 1.77 | 6.07 |
| LSTM-BO | 0.72 | 3.49 | 1.20 | 4.1 |
| BiLSTM | 0.60 | 3.02 | 2.02 | 4.29 |
| BiLSTM-BO | 0.56 | 2.46 | 1.01 | 3.27 |

It is important to note that the testing phase involved forecasting up to 30 days ahead for the daily German dataset and up to 12 months ahead for the monthly Brazilian dataset. This forecasting horizon adds practical relevance to the evaluation, as it reflects realistic operational scenarios for both short-term and long-term electricity demand planning.

For the German dataset, BiLSTM-BO achieves the lowest MAE (13.46 train, 37.59 test) and MAPE (0.99% train, 2.50% test), followed by BiLSTM, LSTM-BO, and LSTM. For the Brazilian dataset, BiLSTM-BO also records the lowest MAE (376,315 train, 1,161,574 test) and MAPE (0.99% train, 2.67% test), with LSTM-BO ranking second. Standard LSTM consistently shows the highest errors in both datasets.

While Tables 4 and 5 provide a clear summary of average accuracy, they do not convey the variability of model performance across folds. Therefore, Figures 8 and 9 present boxplots of MAPE values for the German and Brazilian datasets, respectively, to illustrate the distribution and stability of each model’s performance. These results correspond to models trained using min–max normalization, which was adopted for consistency across experiments after comparative evaluation with z-score standardization.

For the German dataset (Figure 8), BiLSTM-BO achieves the lowest median MAPE and smallest spread in both training and testing phases, indicating the best and most consistent forecasting performance with minimal signs of overfitting.

The other three models show trade-offs between median accuracy and stability, with standard LSTM being the least effective and exhibiting the highest variability in testing.

To assess whether the observed differences in predictive accuracy between LSTM and LSTM-BO were statistically significant, the Kolmogorov–Smirnov Predictive Accuracy (KSPA) test proposed by Hassani and Silva [39] was applied to both fold-level and step-ahead (1–30) error distributions for the German dataset. The results are presented in Tables 8 and 9.

The KSPA is a non-parametric test based on the principles of the Kolmogorov–Smirnov (KS) test and serves two complementary purposes. First, it evaluates whether there is a statistically significant difference between the distributions of forecast errors produced by two competing models. Second, it applies the concept of stochastic dominance to determine whether the model with lower mean errors also exhibits a stochastically smaller error distribution, thereby providing a directional assessment of predictive superiority. The test was implemented using the Hassani. Silva package in R, and both absolute and squared error variants were examined at a 5% significance level.

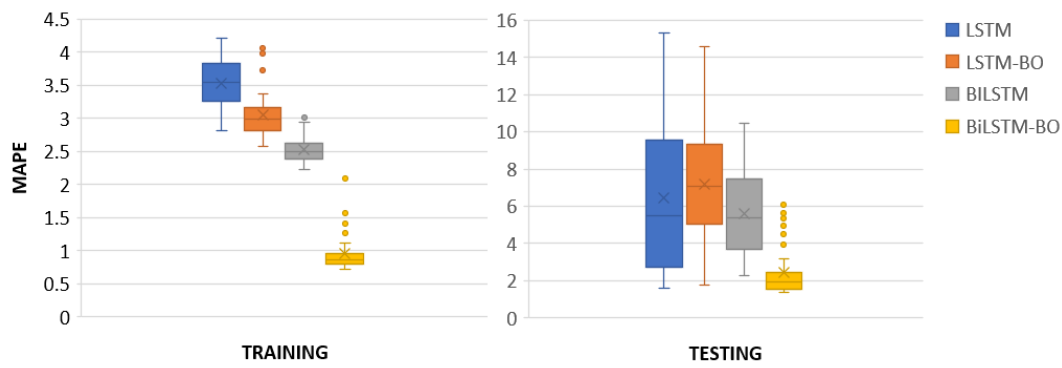


Figure 8. Boxplot for the MAPEs obtained from LSTM and BiLSTM models with and without BO for the daily electricity consumption in Germany

Table 8. Results of the absolute and square error one-sided KSPA test comparing LSTM and LSTM-BO on the German test set across cross-validation folds

| Fold | MAE | | p-value | Significance (Better Model) |
|------|---------------|---------------|---------|--------------------------------|
| | LSTM | LSTM-BO | | |
| 1 | 149.72 | 144.16 | 0.4390 | No |
| 2 | 192.13 | 131.67 | 0.01729 | Yes, (LSTM-BO) |
| 3 | 123.10 | 166.85 | 0.0354 | Yes, (LSTM) |
| 4 | 160.73 | 196.66 | 0.0078 | Yes, (LSTM) |
| 5 | 131.77 | 153.13 | 0.0675 | No |
| 6 | 179.16 | 184.72 | 0.1977 | No |
| 7 | 79.73 | 181.04 | 0.0013 | Yes, (LSTM) |
| 8 | 104.72 | 132.52 | 0.0354 | Yes, (LSTM) |
| 9 | 141.61 | 53.20 | 0.0000 | Yes, (LSTM-BO) |
| 10 | 182.50 | 23.15 | 0.0000 | Yes, (LSTM-BO) |
| 11 | 99.91 | 41.55 | 0.0001 | Yes, (LSTM-BO) |
| 12 | 36.16 | 34.33 | 0.3048 | No |
| 13 | 40.80 | 38.77 | 0.7441 | No |
| 14 | 38.81 | 38.77 | 0.1197 | No |
| 15 | 130.80 | 131.02 | 0.5909 | No |
| 16 | 84.52 | 71.98 | 0.1977 | No |
| 17 | 35.26 | 29.73 | 0.4390 | No |
| 18 | 29.36 | 38.60 | 0.5909 | No |
| 19 | 68.78 | 38.60 | 0.0004 | Yes, (LSTM-BO) |
| 20 | 82.44 | 46.33 | 0.0013 | Yes, (LSTM-BO) |
| 21 | 26.02 | 51.09 | 0.0354 | Yes, (LSTM) |
| 22 | 143.94 | 37.65 | 0.0001 | Yes, (LSTM-BO) |
| 23 | 40.44 | 48.30 | 0.3048 | Yes, (LSTM-BO) |
| 24 | 97.72 | 58.57 | 0.0078 | Yes, (LSTM-BO) |
| 25 | 166.12 | 73.25 | 0.0033 | Yes, (LSTM-BO) |
| 26 | 120.51 | 89.62 | 0.0013 | Yes, (LSTM-BO) |
| 27 | 86.58 | 102.47 | 0.1980 | Yes, (LSTM-BO) |
| 28 | 87.20 | 98.76 | 0.1980 | No |
| 29 | 138.49 | 115.77 | 0.0013 | Yes, (LSTM-BO) |
| 30 | 104.82 | 112.98 | 0.5909 | No |
| 31 | 122.23 | 178.33 | 0.1197 | No |
| 32 | 115.21 | 126.43 | 0.5909 | No |

Note: Values in bold indicate lower MAE, representing better model performance for the corresponding fold.

The results of the absolute and squared error one-sided KSPA tests for each fold and forecast step are presented in Tables 8 and 9, respectively. In these tables, the MAE values are reported for both LSTM and LSTM-BO, with the lower values shown in bold to indicate better performance. As observed in Tables 8 and 9, most comparisons yielded $p > 0.05$, indicating no statistically significant difference in predictive accuracy between the two models. Although the

majority of comparisons were not significant, several folds (Table 8) and medium-range horizons, approximately steps 5 to 6 and 12 to 14 (Table 9), showed significant results favoring LSTM-BO, while none favored the standard LSTM. These findings indicate that the slightly higher average MAPE observed for LSTM-BO on the German test set is not statistically meaningful. Overall, both models achieved comparable accuracy, with the BO-optimized version demonstrating more stable performance across folds and forecasting horizons.

Table 9. Results of the absolute and square error one-sided KSPA test comparing LSTM and LSTM-BO on the German test set across forecasting horizons (1–30 steps ahead)

| Step | MAE | | p-value | Significance (Better Model) |
|------|----------------|---------------|---------|--------------------------------|
| | LSTM | LSTM-BO | | |
| 1 | 50.83 | 50.59 | 0.4620 | No |
| 2 | 61.78 | 50.63 | 0.1368 | No |
| 3 | 64.72 | 55.93 | 0.3282 | No |
| 4 | 71.31 | 61.22 | 0.0801 | No |
| 5 | 77.29 | 54.49 | 0.0047 | Yes, (LSTM-BO) |
| 6 | 78.53 | 55.92 | 0.0438 | Yes, (LSTM-BO) |
| 7 | 77.70 | 52.29 | 0.0801 | No |
| 8 | 89.84 | 68.76 | 0.2188 | No |
| 9 | 91.17 | 73.21 | 0.1368 | No |
| 10 | 76.39 | 68.70 | 0.4620 | No |
| 11 | 87.58 | 78.36 | 0.1368 | No |
| 12 | 97.23 | 75.09 | 0.0438 | Yes, (LSTM-BO) |
| 13 | 110.27 | 74.66 | 0.0224 | Yes, (LSTM-BO) |
| 14 | 102.79 | 68.77 | 0.0107 | Yes, (LSTM-BO) |
| 15 | 101.61 | 75.01 | 0.1368 | No |
| 16 | 96.52 | 90.06 | 0.3282 | No |
| 17 | 83.99 | 80.35 | 0.1368 | No |
| 18 | 97.65 | 91.78 | 0.3282 | No |
| 19 | 109.66 | 87.04 | 0.0801 | No |
| 20 | 118.84 | 92.71 | 0.1368 | No |
| 21 | 115.44 | 86.31 | 0.1368 | No |
| 22 | 113.03 | 91.31 | 0.2188 | No |
| 23 | 118.71 | 115.82 | 0.1368 | No |
| 24 | 107.72 | 117.91 | 0.6105 | No |
| 25 | 130.60 | 140.31 | 0.6105 | No |
| 26 | 140.127 | 144.80 | 0.7578 | No |
| 27 | 145.54 | 159.79 | 0.3282 | No |
| 28 | 163.43 | 158.25 | 0.3282 | No |
| 29 | 172.029 | 171.32 | 0.7578 | No |
| 30 | 180.12 | 192.99 | 0.6105 | No |

Note: Bold values denote the lower MAE, indicating better model performance at each step.

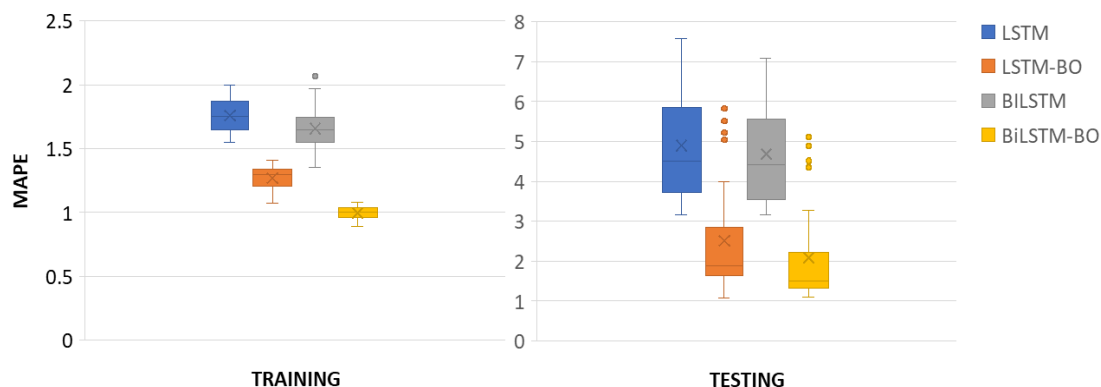


Figure 9. Boxplot for the MAPEs obtained from LSTM and BiLSTM models with and without BO for the monthly electricity consumption in Brazil

Further, for the Brazilian dataset (Figure 9), BiLSTM-BO again demonstrates the lowest median MAPE and smallest spread, followed by LSTM-BO. These results confirm that Bayesian optimization significantly improves both LSTM and BiLSTM performance. Standard LSTM and BiLSTM without optimization perform worse, with LSTM showing the weakest generalization and the highest variability in the test phase.

Overall, applying BO substantially reduced forecasting errors, with the largest gains observed for BiLSTM. For the German dataset, BO reduced BiLSTM's test MAE and MAPE by 53.80% and 54.71%, respectively, compared with the non-optimized BiLSTM baseline, while for Brazil, the reductions were 51.26% and 50.56%. LSTM models also benefited, particularly for Brazil, where test MAE and MAPE decreased by 43.04% and 46.03%, respectively, relative to the non-optimized LSTM baseline. The only exceptions were the German training MAE and test MAPE for LSTM, where BO resulted in slightly higher errors. As shown in Figure 6, the distribution of MAPE values for LSTM-BO in the German dataset exhibits a narrower spread compared with LSTM, suggesting that BO improved the stability of model performance even when the median error was marginally higher. These results highlight the strong positive impact of BO, particularly when combined with the bidirectional architecture.

3.2 Discussion

BO introduces additional computational overhead due to its iterative search and model evaluation process. All experiments were conducted on the Kaggle platform, utilizing a single NVIDIA P100 GPU (16 GB RAM) in a Python environment with TensorFlow and Keras APIs. The training time varied across configurations: for the BO-enhanced models, the runtime ranged from approximately 38 minutes for LSTM (Brazil) to 3 hours 43 minutes for BiLSTM (Germany), whereas the non-optimized models required only 14–20 minutes to train. Although BO substantially increased the computational cost, the optimization was performed offline and only once for each dataset. Once the optimal hyperparameters were identified, retraining the models for operational forecasting required significantly less time. This balance between computational cost and predictive performance underscores the practicality of BO as a data-driven approach that enhances model robustness while reducing the need for manual tuning in future retraining cycles.

Regarding forecasting results, the BiLSTM-BO model consistently outperformed the other approaches across both datasets, achieving the best accuracy and stability during training and testing. Its strong performance stems from the synergy between bidirectional processing, capturing temporal information from both past and future contexts, and the adaptive parameter search enabled by BO, which optimizes model configuration for higher forecasting accuracy.

The benefit of Bayesian optimization is clear in both the daily and monthly experiments. For the German dataset, BiLSTM-BO produced the lowest average errors and the most consistent results across folds, maintaining reliable accuracy even when predicting up to 30 days ahead. For the Brazilian dataset, BiLSTM-BO again achieved the best accuracy, followed closely by LSTM-BO. These outcomes emphasize how effective hyperparameter optimization can strengthen model generalization, particularly for lower-frequency

monthly data with longer forecasting horizons of up to 12 months.

By contrast, standard LSTM and BiLSTM models without optimization exhibited higher errors and greater variability, underscoring the limitations of fixed hyperparameters in capturing complex demand patterns. The performance gap between optimized and non-optimized models was more pronounced in the monthly Brazilian dataset, suggesting that tuning plays a more critical role when dealing with smaller datasets and long-term forecasts, where overfitting or underfitting risks are higher.

These findings align with previous studies reporting that bidirectional architectures can capture richer contextual information [36], and that systematic hyperparameter optimization can significantly enhance forecasting performance [23]. Overall, the consistent superiority of BiLSTM-BO across datasets and forecasting horizons demonstrates its robustness and suitability for practical electricity demand forecasting applications. It is also noteworthy that the forecasting improvements observed in this study align with recent calls for more data-efficient optimization strategies in energy forecasting [20]. By demonstrating that Bayesian optimization substantially reduces error rates with relatively modest computational budgets, our findings provide practical guidance for energy researchers and industry practitioners who often face resource constraints.

In addition to model optimization, normalization techniques were also found to influence forecasting performance. Although detailed numerical results are not presented here, additional experiments were conducted using both min–max normalization and z-score standardization. The outcomes were not entirely consistent across datasets. Despite the upward trend in electricity consumption, models trained with min–max normalization often performed better for the Brazil dataset, whereas standardized data yielded slightly lower errors for Germany. This may appear counterintuitive, as standardization is commonly preferred for trending data to center the distribution. However, LSTM and BiLSTM architectures rely on the compatibility of input values with activation functions such as tanh and sigmoid, which operate most effectively within constrained input ranges. These results suggest that the effect of normalization depends on the temporal characteristics of the data. For the Germany dataset (Figure 4), characterized by strong seasonality and relatively stable amplitude, standardization preserved fluctuations around a consistent mean. In contrast, the Brazil dataset (Figure 5) exhibits a pronounced upward trend and larger magnitude variation, where min–max normalization helped maintain proportional scaling and stabilize gradient updates. Overall, no single normalization method consistently outperformed the other; their effectiveness appears linked to data dynamics and model sensitivity. Consequently, the interaction between normalization choice and temporal structure should be carefully considered when designing deep learning pipelines for forecasting tasks.

From a practical standpoint, these results suggest that combining bidirectional LSTM architectures with Bayesian optimization can provide utility companies and energy planners with more reliable forecasts for both short-term and long-term electricity demand. The ability of BiLSTM-BO to maintain high accuracy and stability across different temporal frequencies means it can be adapted to various operational contexts, from daily grid balancing to annual capacity

planning. This versatility is particularly important in the energy sector, where reliable demand forecasts play a key role in optimizing costs, allocating resources efficiently, and maintaining grid stability. Even so, the real-world applicability of the proposed BO-enhanced LSTM and BiLSTM models should be interpreted with care. The Brazilian dataset, for instance, includes the period affected by the COVID-19 pandemic, during which electricity consumption patterns deviated markedly from historical behavior. These irregularities may limit the accuracy of long-term forecasts (up to 12 months ahead) and highlight the importance of incorporating external variables or adaptive retraining methods in future work. Despite these challenges, the results contribute valuable insights into model generalization and the role of parameter optimization in improving electricity demand forecasts.

For future research, incorporating relevant exogenous variables such as temperature, humidity, and other climate indicators may further enhance forecast accuracy, particularly in regions where electricity demand is strongly shaped by weather conditions. Integrating these variables into the BiLSTM-BO framework could lead to more adaptive and robust forecasting models for real-world energy management.

4. CONCLUSION

This study evaluated the performance of LSTM and BiLSTM models, both with and without Bayesian optimization, using electricity consumption data from Germany (daily) and Brazil (monthly). Applying a rolling-window time series cross-validation, the analysis showed that BiLSTM-BO achieved the most accurate and stable forecasts across both datasets, up to 30 days ahead for Germany and up to 12 months ahead for Brazil. The use of Bayesian optimization notably enhanced model generalization, particularly for the lower-frequency monthly data, where the improvement over non-optimized models was more evident.

The findings highlight the advantages of combining bidirectional processing with data-driven hyperparameter tuning, offering a robust and transferable forecasting framework that can support both short-term operational planning and long-term capacity management in the energy sector. From a practical perspective, the adaptability of BiLSTM-BO across different temporal frequencies provides valuable guidance for utilities and energy planners seeking reliable forecasting solutions under diverse operating conditions.

This study is limited to univariate electricity consumption data without incorporating exogenous factors such as weather and socio-economic indicators, or anomalous events like the COVID-19 pandemic. Future research should expand the framework by incorporating climate and external drivers, exploring multi-step forecasting strategies, and testing hybrid approaches that combine deep learning with statistical or physics-informed approaches. Such advancements could further enhance the adaptability and resilience of forecasting systems, supporting the development of sustainable and data-driven energy management practices.

ACKNOWLEDGMENT

This research was funded by the Directorate of Research

and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology under Contract No.: 105/C3/DT.05.00/PL/2025 and Assignment Agreement Letter Number from LPPM UNS: 1186.1/UN27.22/PT.01.03/2025.

REFERENCES

- [1] Lee, Y.W., Tay, K.G., Choy, Y.Y. (2018). Forecasting electricity consumption using time series model. *International Journal of Engineering & Technology*, 7(4.30): 218-223.
- [2] Elsaraiti, M., Ali, G., Musbah, H., Merabet, A., Little, T. (2021). Time series analysis of electricity consumption forecasting using ARIMA model. In 2021 IEEE Green Technologies Conference (GreenTech), Denver, CO, USA, pp. 259-262. <https://doi.org/10.1109/GreenTech48523.2021.00049>
- [3] Shiwakoti, R.K., Charoenlarnnoppa, C., Chapagain, K. (2023). Time series analysis of electricity demand forecasting using seasonal arima and an exponential smoothing model. In 2023 International Conference on Power and Renewable Energy Engineering (PREE), Tokyo, Japan, pp. 131-137. <https://doi.org/10.1109/PREE57903.2023.10370319>
- [4] Nugraha, R.I. (2024). Time series analysis for electricity demand forecasting: A comparative study of ARIMA and exponential smoothing models in Indonesia. *Information Technology International Journal*, 2(2): 78-88. <https://doi.org/10.33005/itij.v2i2.23>
- [5] Bashir, T., Haoyong, C., Tahir, M.F., Liqiang, Z. (2022). Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN. *Energy Reports*, 8: 1678-1686. <https://doi.org/10.1016/j.egyr.2021.12.067>
- [6] Arslan, S. (2022). A hybrid forecasting model using LSTM and Prophet for energy consumption with decomposition of time series data. *PeerJ Computer Science*, 8: e1001. <https://doi.org/10.7717/peerj-cs.1001>
- [7] Torres, J.F., Martínez-Álvarez, F., Troncoso, A. (2022). A deep LSTM network for the Spanish electricity consumption forecasting. *Neural Computing and Applications*, 34(13): 10533-10545. <https://doi.org/10.1007/s00521-021-06773-2>
- [8] Albahli, S. (2025). LSTM vs. Prophet: Achieving superior accuracy in dynamic electricity demand forecasting. *Energies*, 18(2): 278. <https://doi.org/10.3390/en18020278>
- [9] Sulandari, W., Yudhanto, Y., Zukhronah, E., Slamet, I., Pardede, H.F., Rodrigues, P.C., Lee, M.H. (2025). Hybrid Prophet-NAR model for short-term electricity load forecasting. *IEEE Access*, 13: 7637-7649. <https://doi.org/10.1109/ACCESS.2025.3526735>
- [10] Samson, T.K., Aweda, F.O. (2024). Seasonal autoregressive integrated moving average modelling and forecasting of monthly rainfall in selected African stations. *Mathematical Modelling of Engineering Problems*, 11(1): 159-168. <https://doi.org/10.18280/mmep.110117>
- [11] Kazeem, R.A., Petinrin, M.O., Akhigbe, P.O., Jen, T.C., Akinlabi, E.T., Akinlabi, S.A., Ikumapayi, O.M. (2023). Forecast of the trend in sales data of a confectionery baking industry using exponential smoothing and

- moving average models. *Mathematical Modelling of Engineering Problems*, 10(1): 1-13. <https://doi.org/10.18280/mmep.100101>
- [12] Sulandari, W., Yudhanto, Y., Subanti, S., Setiawan, C.D., Hapsari, R., Rodrigues, P.C. (2023). Comparing the simple to complex automatic methods with the ensemble approach in forecasting electrical time series data. *Energies*, 16(22): 7495. <https://doi.org/10.3390/en16227495>
- [13] Peng, L., Wang, L., Xia, D., Gao, Q. (2022). Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. *Energy*, 238: 121756. <https://doi.org/10.1016/j.energy.2021.121756>
- [14] Petroșanu, D.M., Pirjan, A. (2020). Electricity consumption forecasting based on a bidirectional long-short-term memory artificial neural network. *Sustainability*, 13(1): 104. <https://doi.org/10.3390/su13010104>
- [15] Chauhan, N., Bansal, S., Malik, M. (2024). BiLSTM model for electricity consumption prediction: A comparative study with traditional machine learning approaches. In 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, pp. 1-6. <https://doi.org/10.1109/DELCON64804.2024.10866016>
- [16] Alizadegan, H., Rashidi Malki, B., Radmehr, A., Karimi, H., Ilani, M.A. (2025). Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction. *Energy Exploration & Exploitation*, 43(1): 281-301. <https://doi.org/10.1177/01445987241269496>
- [17] Van Houdt, G., Mosquera, C., Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8): 5929-5955. <https://doi.org/10.1007/S10462-020-09838-1>
- [18] Snoek, J., Larochelle, H., Adams, R.P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25.
- [19] Yang, T., Li, B., Xun, Q. (2019). LSTM-attention-embedding model-based day-ahead prediction of photovoltaic power output using Bayesian optimization. *IEEE Access*, 7: 171471-171484. <https://doi.org/10.1109/ACCESS.2019.2954290>
- [20] Bischl, B., Binder, M., Lang, M., Pielok, T., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2): e1484. <https://doi.org/10.1002/widm.1484>
- [21] Alibrahim, H., Ludwig, S.A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In 2021 IEEE Congress on Evolutionary Computation (CEC), Kraków, Poland, pp. 1551-1559. <https://doi.org/10.1109/CEC45853.2021.9504761>
- [22] Habtemariam, E.T., Kekeba, K., Martínez-Ballesteros, M., Martínez-Álvarez, F. (2023). A Bayesian optimization-based LSTM model for wind power forecasting in the Adama district, Ethiopia. *Energies*, 16(5): 2317. <https://doi.org/10.3390/en16052317>
- [23] Michael, N.E., Hasan, S., Al-Durra, A., Mishra, M. (2022). Short-term solar irradiance forecasting based on a novel Bayesian optimized deep Long Short-Term Memory neural network. *Applied Energy*, 324: 119727. <https://doi.org/10.1016/j.apenergy.2022.119727>
- [24] Herrera-Casanova, R., Conde, A., Santos-Pérez, C. (2024). Hour-ahead photovoltaic power prediction combining BiLSTM and Bayesian optimization algorithm, with bootstrap resampling for interval predictions. *Sensors*, 24(3): 882. <https://doi.org/10.3390/s24030882>
- [25] Marweni, M., Yahyaoui, Z., Chaabani, S., Hajji, M., Mansouri, M., Bouazzi, Y., Mimouni, M.F. (2025). Forecasting of solar irradiance and power in uncertain photovoltaic systems using BiLSTM and Bayesian optimization. *Arabian Journal for Science and Engineering*, 50(14): 10763-10777. <https://doi.org/10.1007/s13369-024-09818-5>
- [26] Wang, S., Ma, C., Gao, H., Deng, D., Fernandez, C., Blaabjerg, F. (2025). Improved hyperparameter Bayesian optimization-bidirectional long short-term memory optimization for high-precision battery state of charge estimation. *Energy*, 328: 136598. <https://doi.org/10.1016/j.energy.2025.136598>
- [27] Li, H., Lin, Z., An, Z., Zuo, S., et al. (2022). Automatic electrocardiogram detection and classification using bidirectional long short-term memory network improved by Bayesian optimization. *Biomedical Signal Processing and Control*, 73: 103424. <https://doi.org/10.1016/j.bspc.2021.103424>
- [28] Germany electricity power for 2006-2017. <https://www.kaggle.com/datasets/mvianna10/germany-electricity-power-for-20062017>.
- [29] Power statistics. <https://www.entsoe.eu/data/power-stats/>.
- [30] da Silva, L.C.P., da Silva Cordeiro, J., da Costa, K., Saboya, N., Rodrigues, P.C., López-Gonzales, J.L. (2025). Time series forecasting via integrating a filtering method: An application to electricity consumption. *Computational Statistics*, 1-20. <https://doi.org/10.1007/s00180-024-01595-x>
- [31] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8): 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [32] Yu, Y., Si, X., Hu, C., Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31(7): 1235-1270. https://doi.org/10.1162/neco_a_01199
- [33] Bilgili, M., Pinar, E. (2023). Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye. *Energy*, 284: 128575. <https://doi.org/10.1016/j.energy.2023.128575>
- [34] Schuster, M., Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673-2681. <https://doi.org/10.1109/78.650093>
- [35] Graves, A., Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6): 602-610. <https://doi.org/10.1016/j.neunet.2005.06.042>
- [36] García, F., Guijarro, F., Oliver, J., Tamošiūnienė, R. (2024). Foreign exchange forecasting models: LSTM and BiLSTM comparison. *Engineering Proceedings*, 68(1): 19. <https://doi.org/10.3390/engproc2024068019>
- [37] Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series

- forecasting. *International Journal of Forecasting*, 36(1): 75-85. <https://doi.org/10.1016/j.ijforecast.2019.03.017>
- [38] Sulandari, W., Subanar, S., Lee, M.H., Rodrigues, P.C. (2020). Time series forecasting using singular spectrum analysis, fuzzy systems and neural networks. *MethodsX*, 7: 101015. <https://doi.org/10.1016/j.mex.2020.101015>
- [39] Hassani, H., Silva, E.S. (2015). A Kolmogorov-Smirnov based test for comparing the predictive accuracy of two sets of forecasts. *Econometrics*, 3(3): 590-609. <https://doi.org/10.3390/econometrics3030590>