Load Forecasting for an Industrial Building with Aggregated and Disaggregated Load

Daniela Eser*, Michael Suriyah, Thomas Leibfried
Institute of Electric Energy Systems and High-Voltage Technology (IEH)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
*daniela.eser@kit.edu

***Abstract*—Short-term load forecasting in multipurpose research buildings is challenging because diverse load categories exhibit heterogeneous and often unpredictable usage patterns, complicating the design of reliable energy-management strategies. This work provides a systematic empirical benchmark that compares direct total load prediction with a disaggregated approach that forecasts each load category separately and aggregates the results. Using Long Short-Term Memory networks applied to different load categories across multiple training durations and seasons, we deliver practical guidance on when to use each aggregation strategy. Our analysis reveals that direct forecasting achieves lower mean absolute error in most scenarios, with disaggregation providing benefits only for loads with independent, sporadic usage patterns. We explore conditions favoring disaggregated predictions, including behavioral diversity between categories, predictability of individual components, and representative training data.**

***Index Terms*—demand forecasting, short term load forecasting, aggregation level, time series analysis, building energy consumption**

I. Introduction

The electrical grid is transforming as renewable energy, electric vehicles, and heat pumps reduce the significance of traditional load profiles. While grid-scale forecasting is well-established, there is growing demand for localized forecasts in microgrids and smart buildings to better control distributed resources [1]. Recent research has focused on improving these short-term forecasts, but the limited availability of quality training data for modern algorithms makes this harder [2].

One challenge in forecasting concerns aggregation levels: should diverse loads be predicted as an aggregate or forecasted separately and then combined? Recent studies on prosumer systems [3]–[5] suggest that disaggregated forecasting can outperform direct aggregate prediction, though whether this extends to different load types remains unclear.

This paper investigates whether category-specific forecasting outperforms direct total load prediction in a real multipurpose building. Both methods are compared under different seasonal and time scales by applying Long Short-Term Memory (LSTM) models to different load types. This contribution provides suggestions when each method is most appropriate while highlighting limitations of established methods when working with realistic data.

II. Related Work

Short-term load forecasting (STLF) has been extensively studied in multiple directions. This section reviews the relevant literature that motivates the investigation of disaggregated forecasting strategies for multipurpose buildings.

1) Forecasting Algorithms: Numerous studies have compared forecasting techniques for building and grid-level loads. Comprehensive reviews of common approaches [6], [7] conclude that each method has unique strengths and limitations, making the optimal choice highly application-dependent. In [8], the authors further demonstrate that different load types correspond better to different forecasting models, suggesting that using the same approach may be suboptimal for heterogeneous load groups. Among the various methods reviewed in literature [1], [2], [7]–[9] LSTM networks have emerged as one of the most widely adopted approaches for short-term building load forecasting, with demonstrated success across diverse applications [3], [5], [10]–[12]. Their effectiveness stems from the ability to capture long-term temporal dependencies and learn complex non-linear patterns in sequential data [10], making them particularly suitable for building loads with irregular occupancy and usage patterns across various scales, from single household profiles [11] to commercial and mixed-use buildings. Since our research focuses on categorical decomposition strategies rather than algorithm optimization, we employ LSTM as a representative and well-established forecasting method.

2) Aggregated and Disaggregated Forecasting: Traditionally, aggregated loads at the grid or substation level have been the primary focus of forecasting research. Aggregation tends to smooth individual consumption volatility, as random behavioral variations cancel out across many consumers, producing more predictable patterns [9]. Recent research challenges the assumption that aggregation always improves forecasts. Several studies have demonstrated that separately forecasting photovoltaic (PV) generation and load demand, then combining the predictions, yields higher accuracy compared to directly forecasting the total load [3]–[5]. This improvement occurs because generation (driven primarily by weather) and load (driven by occupancy and usage patterns) follow fundamentally different dynamics.

In [13], this concept is extended to industrial production

979-8-3315-8138-1

forecasting, comparing aggregated, disaggregated, and combined approaches. Their results suggest that the best strategy depends on the similarity of underlying components and the ability of models to exploit component-specific patterns. This principle was formalized in [14] by quantifying correlations between input features and load. They showed that additional information such as weather or day-of-week only improves predictions when correlation exists; otherwise, additional features degrade performance. This finding supports the notion that forecasting should group time series with similar driving factors while separating those with distinct influences.

3) Research Gap: While prior work demonstrates benefits of separating generation from load in prosumer forecasting, and recognizes that different building types exhibit distinct patterns, limited research examines whether disaggregation benefits extend to heterogeneous loads within a single multipurpose building. Specifically, it remains unclear whether such loads should be forecasted in categories based on behavioral similarity and aggregated afterwards to improve forecast accuracy compared to direct total load prediction.

By using real operational data from an active research facility, this work also addresses practical forecasting challenges including missing data, irregular usage patterns, and atypical load characteristics.

III. Methodology

This paper tests the hypothesis that disaggregated forecasting—training separate models for distinct load categories and summing their predictions—will outperform direct forecasting of total building load, provided two conditions: (1) loads within each category must share similar temporal patterns, enabling category-specific models to learn coherent behaviors, while (2) different categories must exhibit sufficiently distinct patterns that a single model cannot capture all behaviors simultaneously. This approach is used for predicting the total building load as well as the total category load of the introduced categories. For predicting the category loads, intercategory diversity is limited since subloads were clustered by similarity, so aggregation effects should dominate. However, for total building load prediction, both key conditions should be met. If the hypothesis holds, disaggregated forecasting should outperform direct prediction.

The remainder of this section describes the dataset, load categorization, LSTM configuration, and experimental design used to compare the direct and disaggregated forecasting strategies in a multipurpose building.

A. Dataset Description

The dataset originates from an institute building currently being transformed into a smart energy complex, as described in [15] and [16]. The building serves multiple purposes including teaching facilities, experimental laboratories, industrial processes, and office spaces. The electrical infrastructure connects to the medium-voltage grid through four transformers with 1.3 MVA combined capacity. This study focuses on active power measurements for phase 1 from 35 feeder-level loads at

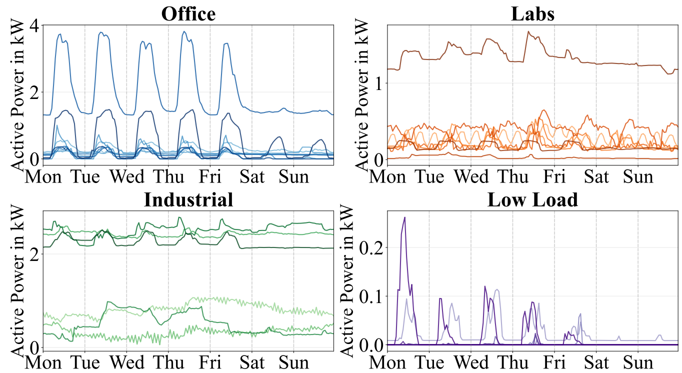

Fig. 1. Avarage weekly load profiles categorized by load type.

200 ms resolution, collected since February 2024 with 98.15% data coverage. These loads exhibit varying dependencies on workday schedules and weather conditions, with varying baseline consumption levels.

1) Load Categorization: Based on preliminary correlation analysis using the hierarchical clustering of *seaborn* [17], the 35 feeder-load profiles were grouped into four distinct categories according to their temporal patterns and operational characteristics. Fig. 1 illustrates the average weekly profile for each category over a one-year period.

- **Office** (9 loads): Distinct daytime peaks during workdays from ceiling lighting, wall outlets (computers/printers), and lecture hall air conditioning, indicating strong occupancy dependence with high mean consumption.
- **Labs** (8 loads): Less predictable patterns due to volatile experimental activities compared to routine office operations. Includes one high-demand profile.
- **Industrial** (6 loads): Air compressors, laboratory air conditioning, and climate chambers demonstrating more constant operation with reduced occupancy dependence. Half exhibit high demand.
- **Low Load** (12 loads): Mean load below 25 W from out-of-order or rarely utilized assets (e.g., heavy-duty crane used briefly).

2) Dataset Challenges: The multipurpose use of the building, combining diverse load types with different behavioral patterns, complicates total load prediction. The total load over 1.5 years is visualized in Fig. 2 with test periods highlighted. The building exhibits substantial variability, with hourly load ranging from approximately 5 kW to peaks exceeding 43 kW, without clear seasonal patterns. As a research institution, the building lacks controlled working hours typical of manufacturing facilities, resulting in irregular occupancy patterns. Decades-old infrastructure (non-LED lighting) and experimental power loops (where feeders exchange power internally) create unexpected consumption patterns and occasional negative power flows for individual feeders.

This study's findings are limited to one building and should be interpreted as a case study requiring validation across diverse building types and operational contexts.

979-8-3315-8138-1

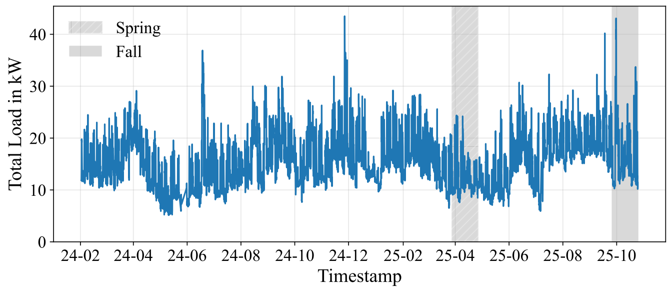

Fig. 2. Total load consumption over the entire dataset with test periods highlighted.

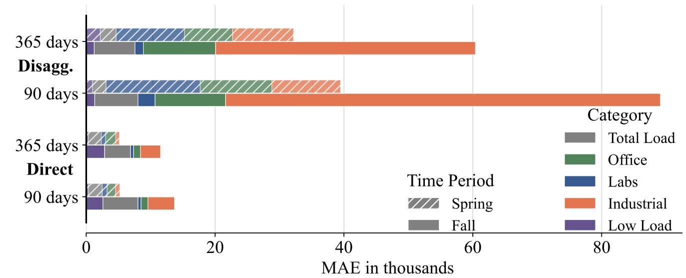

Fig. 3. MAE results for all 40 forecasting scenarios.

B. Forecasting Model Configuration

Forecasting results are influenced by various factors. To isolate the effect of aggregation methods on performance, modeling choices were kept constant and straightforward across all experiments. This section gives an overview on those parameters.

1) Model: A 2-layer encoder-decoder LSTM with 128 hidden states based on the Nixtla repository [18] was used, drawing on standard architecture as a benchmark approach. Training is limited to a maximum of 2000 steps (adjusted to the used horizon) with early stopping enabled to prevent overfitting.

2) Training Configuration: A 24-hour day-ahead forecasting horizon with an hourly resolution aligns with the intended application of controlling the building's Hybrid Energy Storage System (HESS) for optimal energy management.

Two training durations were tested: 90 days and 365 days of historical data. The 90-day training window was selected based on literature [12] demonstrating sufficient performance with this duration, while the 365-day window allows the model to capture full seasonal variations. The training durations are limited by data availability, as the measurement system has been operational for approximately 1.5 years. Rather than substituting missing values, the LSTM implementation processes only timestamps with available data. This approach avoids introducing artificial patterns that could degrade model performance. With 98.15% data coverage, the impact of missing data on training and testing is minimal.

The model operates in univariate mode, using only historical load data for a single time series as input and neglecting interactions between time series. As established in [14], exogenous variables can improve forecasting when correlations exist. However, preliminary analysis showed that exogenous features interact complexly with the aggregation approach: sometimes improving aggregated totals while worsening individual category forecasts. To isolate aggregation structure from feature-selection dependencies, this benchmark excludes exogenous variables, establishing a controlled baseline while enabling future application of category-specific exogenous inputs.

C. Experimental Procedure

Preliminary experiments revealed high sensitivity when evaluating single-day forecasts, with performance strongly dependent on whether the test day exhibited patterns similar to the training period. To obtain robust results while working within data availability constraints (approx. 1.5 years), each experimental configuration was evaluated over 30 consecutive test days. This provides sufficient data to assess typical performance while capturing day-to-day variability.

The rolling training approach updates the training window daily, using the 90 or 365 days immediately prior to each prediction day. This approach ensures the model always trains on the most recent data and prevents error accumulation across consecutive predictions. To ensure that results reflect genuine differences between forecasting strategies rather than seasonal snapshots, tests were conducted during two distinct periods: spring and fall (see Fig. 2). These seasons provide contrasting operational conditions, chosen as the best alternative since the winter period lacked sufficient data for 365-day training.

Mean absolute error (MAE) serves as the primary evaluation metric. MAE provides an easily interpretable measure of average prediction error in the same units as the load, making it straightforward to assess practical forecasting accuracy. Additionally, LSTM models minimize the MAE during training, ensuring consistency between optimization and evaluation. While other metrics like root mean squared error (RMSE) penalize large errors more heavily or provide scale-independent comparisons like mean absolute percentage error (MAPE), MAE is sufficient for this comparative study as the relative ranking of strategies remains consistent across all three metrics.

The following scenarios are examined:

- Forecasting strategy: Direct vs. Disaggregated
- Training duration: 90 days vs. 365 days
- Test season: Spring vs. Fall

IV. Results and Discussion

This section presents the evaluation of direct versus disaggregated forecasting strategies across multiple scenarios.

A. Overall Performance Comparison

The explored scenarios vary across two training durations (90 and 365 days), two test seasons (spring and fall), five aggregation levels (total load and four load categories), and two forecasting strategies (direct and disaggregated prediction). Fig. 3 summarizes the MAE for all 40 experimental scenarios.

979-8-3315-8138-1

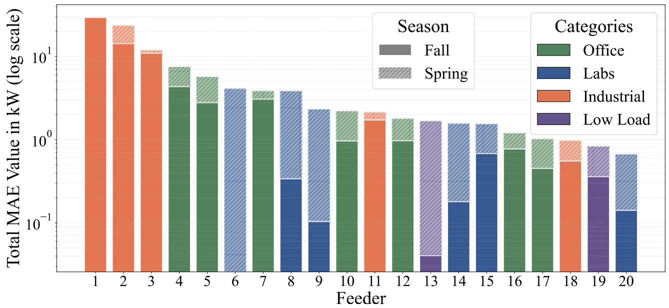

Fig. 4. MAE of feeder-level forecasts (used in disaggregation approach) for 365-day training period, with feeders ranked by error.

Longer training generally improves accuracy, with 365-day training consistently outperforming 90-day training by capturing more operational patterns. However, 365-day training produces more than twice the error for Low Load spring predictions. The characteristic spring behavior appears in the 90-day window but not in the remaining 275 days. This highlights a known challenge in time series forecasting: training data must represent the test period, regardless of training duration.

Forecast accuracy varies between spring and fall test periods. Most categories show consistently lower errors in one season, suggesting that some load patterns are inherently more predictable during certain times of year. The seasonal variation in forecast performance underscores the importance of multi-season evaluation to ensure model robustness.

B. Category Level

The stacked bars in Fig. 3 reveal that different load categories contribute with varying error margins. The office and industrial categories dominate the error distribution, while labs and Low Load categories contribute less to the total error. This pattern reflects both the magnitude of loads in each category and their relative predictability.

Comparing aggregation methods at the category level, direct prediction mostly achieves lower MAE than the disaggregated approach, with the Low Load category presenting a notable exception where disaggregation yields superior predictions during the fall test period. The Low Load category's positive response to disaggregated forecasting is particularly noteworthy. The sporadic, independent usage patterns of these loads mean that aggregation does not produce the smoothing effect typically expected. In this case, forecasting each load's simple on-off pattern separately proves more effective than modeling their combined behavior, resulting in a 43% MAE reduction compared to direct prediction.

However, this improvement is small compared to the risks of disaggregation. When direct prediction excels, the biggest decline is a 20-fold increase in MAE for the lab category. This contrast indicates that while disaggregation rarely provides substantial benefits at category level, it carries significant risk of performance degradation.

C. Load Level

To identify which loads drive category-level performance differences, Fig. 4 presents the 20 highest MAE values for individual loads in the disaggregated approach with 365-day training. The logarithmic scale reveals that a small subset of highly problematic loads dominates the error distribution. The industrial and office categories account for the top 5 high-error loads, explaining their large degradation under disaggregated forecasting, while the Low Load category contributes minimal error due to low absolute consumption. Test period significantly influences individual load accuracy. Spring test patterns resembled training data for industrial and office loads, whereas fall data contained outliers and atypical usage patterns not well-represented in either training window.

The climate chamber (industrial) emerges as the most problematic load (MAE > 28 kW), operating infrequently (twice in 365 days) with high peaks (16 kW). During spring, it remained inactive, enabling accurate predictions, whereas in fall two activations caused large errors. Two experimental sub-panels also contribute significantly to industrial category error where inconsistent schedules and power loops (see Section III-A) further complicate industrial predictions.

Office loads show day-to-day variability in their characteristic bell-curve profiles, with fall capturing more variations.

Almost all lab loads appear in the top 20 errors, confirming that experimental activities—with unpredictable type, duration, and intensity—are fundamentally difficult to forecast from historical data alone.

The individual load analysis reveals that certain load types are fundamentally difficult to predict regardless of method. When forecasted individually, their errors directly impact summed category predictions.

D. Total Load Prediction

Contrary to the initial hypothesis, direct forecasting of the total building load outperforms the disaggregated approach in nearly all scenarios. Only one configuration—90-day training in spring—shows marginally lower MAE ($< 1\%$) for the disaggregated approach, though this difference is not substantial enough to recommend the method given its significantly higher computational cost (four times higher) of training four separate models.

Fig. 5 presents the spring test period predictions for both forecasting strategies using 365-day training. Both forecasting methods successfully capture the general daily load curve, capturing the characteristic occupancy-driven patterns. However, both methods also produce over- and under-predictions at various time points, with no single method demonstrating consistent superiority across all time steps.

A notable observation is that the two forecasting strategies show complementary error characteristics. When the direct prediction method produces substantial errors at certain time points, the disaggregated method often achieves predictions closer to the actual load, and vice versa. The complementary nature of their errors indicates that neither method captures all

979-8-3315-8138-1

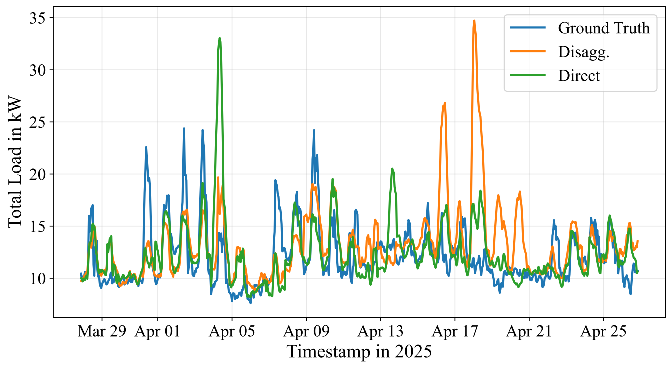

Fig. 5. Load forecast for the spring test period using 365 days of training.

relevant patterns on its own, and their strengths lie in different aspects of the load profile.

V. Conclusion

This study evaluated whether disaggregated forecasting of correlated load groups could improve prediction accuracy for a multipurpose institute building. The central hypothesis—that forecasting load categories separately would outperform direct total load forecasting—was not supported. For most scenarios, direct prediction achieved lower errors. The key difference from successful prosumer forecasting lies in component characteristics: PV generation and load profiles are fundamentally different, whereas load categories in this study may not be sufficiently distinct. Moreover, a third critical condition emerged: individual loads must be inherently predictable from historical data. When loads exhibit high volatility due to rare events, unscheduled operations, or irregular patterns, disaggregation amplifies rather than mitigates forecasting challenges. The building's multipurpose use varies heavily due to its research and teaching purpose, placing an upper bound on achievable accuracy regardless of method. This study was somewhat limited by available data (approximately 1.5 years), so further investigations could lead to different results.

Nevertheless, when analyzing the predicted total load over time, the disaggregated approach shows promise in handling some time steps better than direct forecasting, with results depending highly on whether the training period captured similar behavior. Mixed approaches combining direct and disaggregated strategies using ensemble methods are particularly promising because of the observed complementary error patterns. The disaggregated approach's potential to apply category-specific exogenous features represents an important avenue for future research. Our preliminary analysis revealed multifaceted interactions between exogenous variables, aggregation strategy, and seasonal context. Future work should explore these relationships across diverse building types, load patterns, and temporal contexts to fully exploit disaggregation's flexibility in feature selection.

While based on a single building, the methodology's reliance on having multiple distinct load categories suggests potential applicability to similar building types or larger campus complexes. Buildings with a more uniform load profile (e.g. pure office spaces) are likely to benefit less from disaggregation. However, validation across diverse operational contexts remains necessary.

Similarly, these findings are based solely on LSTM networks. Validation with other forecasting architectures would strengthen these conclusions.

In summary, while direct forecasting remains recommended for multipurpose buildings with diverse loads, this work identifies potential for successful disaggregation and proposes hybrid ensemble strategies that combine the strengths of both approaches.

References

[1] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar, R. Ahmad, and A. H. Vakili, "State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques," *Electr. Power Syst. Res.*, vol. 225, p. 109792, 2023.

[2] S. Akhtar *et al.*, "Short-term load forecasting models: A review of challenges, progress, and the road ahead," *Energies*, vol. 16, no. 10, 2023.

[3] A. Ndife, D. Steen, and L. Tuan, "Scalable and lightweight machine learning based load forecast: Netload versus disaggregrated forecast," in *2024 IEEE PES ISGT EUROPE*, 10 2024, pp. 1–5.

[4] S. Kloibhofer, M. Stifter, F. Leimgruber, and B. V. Rao, "Comparing and improving residential demand forecast by disaggregation of load and pv generation," *CIRED Open Access Proc. J.*, vol. 2017, pp. 1638–1641, 10 2017.

[5] A. Stratman, T. Hong, M. Yi, and D. Zhao, "Net load forecasting with disaggregated behind-the-meter pv generation," in *2022 IEEE Industry Applications Society Annual Meeting (IAS)*, 2022, pp. 1–5.

[6] D. Upadhaya, R. Thakur, and N. K. Singh, "A systematic review on the methods of short term load forecasting," in *2019 2nd Int. Conf. Power Energy, Environment and Intelligent Control (PEEIC)*, 2019, pp. 6–11.

[7] V. Srivastava, S. Kumar, K. Hazarika, U. Varshney, and S. Srivastava, "Review on electrical load forecasting approaches," in *2024 Intelligent Systems and Machine Learning Conference (ISML)*, 2024, pp. 710–715.

[8] J. Wen *et al.*, "A review of research on power system load forecasting methods," in *2024 5th International Conference on Smart Grid and Energy Engineering (SGEE)*, 2024, pp. 55–61.

[9] M. Jacob, C. Neves, and D. Vukadinović Greetham, *Short Term Load Forecasting*. Cham: Springer International Publishing, 2020, pp. 15–37.

[10] S. Muzaffar and A. Afshari, "Short-term load forecasts using lstm networks," *Energy Procedia*, vol. 158, pp. 2922–2927, 2019.

[11] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[12] Z. Chen *et al.*, "Load forecasting based on lstm neural network and applicable to loads of "replacement of coal with electricity"," *Journal of Electrical Engineering & Technology*, vol. 16, 04 2021.

[13] K. Park, S. Yoon, and E. Hwang, "Hybrid load forecasting for mixed-use complex based on the characteristic load decomposition by pilot signals," *IEEE Access*, vol. 7, pp. 12 297–12 306, 2019.

[14] X. Tang, H. Chen, W. Xiang, J. Yang, and M. Zou, "Short-term load forecasting using channel and temporal attention based temporal convolutional network," *Elect. Power Syst. Res.*, vol. 205, p. 107761, 2022.

[15] F. Gielnik, J. Geis-Schroer *et al.*, "Establishing a power hardware-in-the-loop environment with a smart energy complex," in *2022 57th Int. Universities Power Engineering Conference (UPEC)*. IEEE, 2022.

[16] J. Geis-Schroer, D. Eser *et al.*, "Establishing a switchable experimental power grid in the distribution system of a real building," in *2023 8th IEEE Workshop on the Electronic Grid (eGRID)*. IEEE, 2023.

[17] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[18] K. G. Olivares, C. Challú, F. Garza, M. M. Canseco, and A. Dubrawski, "NeuralForecast: User friendly state-of-the-art neural forecasting models." PyCon Salt Lake City, Utah, US 2022, 2022.

979-8-3315-8138-1