Multi-Source Feature Integration for Short-Term Load Forecasting in Off-Grid Residential Microgrids Using Stacked LSTM Networks

Kathi Avinash
Department of EEE
Vignan's Institute of Information Technology,
Duvvada, Visakhapatnam, Pin: 530049
Andhra Pradesh, India
avinash.k@vignaniit.edu.in

Venkata Akhilesh Ranga Reddy
Sr. Application Architect,
Health Care Service Corporation,
Richardson, Texas, United States
venkataakhileshkumar@gmail.com
ORCID ID:0009-0008-4140-2299

D Manohari
Department of Computer Science & Engineering (CSE)
PERI Institute of Technology,
Chennai – 600048, India
pericseofficial@gmail.com

Dr. Sriraman Kothuri
Associate Professor, Department of CSE
Alliance School of Advanced Computing,
Alliance University
Anekal, Bengaluru-562106, Karnataka
sriramankothuri@gmail.com

R Geetha
Department of Artificial Intelligence and Machine Learning Cambridge Institute of Technology
KR Puram, Bengaluru-560036, India.
geetha.aiml@cambridge.edu.in

B Swarna
University Centre for Research & Development, Chandigarh University,
Mohali 140413, India.
Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences, Chennai, India
bswarna261@gmail.com

***Abstract*—Short-term load forecasting is critical for demand–response scheduling and renewable dispatch in smart city microgrids. This paper proposes a stacked Long Short-Term Memory (LSTM) load forecasting framework applied to the Mafate off-grid residential microgrid dataset from Reunion Island, France. Electricity demand from three households is aggregated at hourly resolution and combined with a 15-dimensional feature vector spanning global horizontal irradiance, ambient temperature, relative humidity, solar PV output, cyclic hour encodings, and autoregressive load lags. A two-layer LSTM with a 24-hour look-back window is benchmarked against GRU, vanilla RNN, naive persistence, and a 24-hour moving average across 5,460 hourly observations (May–December 2020). The LSTM achieves RMSE = 0.0856 kW and MAPE = 20.94% — reductions of 20.1% and 30.6% over naive persistence. All three deep learning architectures decisively outperform statistical baselines. Feature correlation analysis identifies the 1-hour load lag (r = 0.681) and solar plant output (r = 0.737) as the strongest demand predictors.**

I. INTRODUCTION

Smart city microgrids confront two interweaving challenges; the accelerating demand due to multiplicative proliferation of buildings, EV and IoT devices combined with solar PV integration with residential storage producing a non-stationary, weather-coupled load signal that classical statistical forecasters are not able to track reliably [1]. Real-time demand–response, peak-period grid stress reduction, and economic dispatch of renewable resources [2], among others, all depend on accurate short-term load forecasting. ARIMA and exponential smoothing algorithms belong to classical methods that restrict [3], as they assume linearity and stationarity. These thus worsen under renewable-driven distributional shift [4], notwithstanding the improvement offered by shallow machine learning methods. Due to the vanishing-gradient limitation in vanilla RNNs, LSTM networks have become the state-of-the-art architecture for sequential load prediction, employing gating mechanisms that enable temporal context preservation throughout the 24-hour intraday cycle (5), (6). The GRU networks give competitive accuracy for a given computational budget [7]. Despite this, three gaps still exist. The majority of benchmarking studies focus on single-household demand, thus missing community scale aggregation effects [8]. However, only a few works have unified the meteorological covariates — irradiance, temperature, humidity — with autoregressive lag features and cyclic time encodings in one input space [9]. And no rigorous comparison of deep learning architectures against naive baselines on real off-grid island microgrids — where grid import will never be possible and forecasting errors map unambiguously to energy poverty, does exist [10]. A stacked two-layer LSTM framework is used to tackle in this paper these three gaps applied on the Mafate off-grid microgrid on Reunion Island [11]. This is done at an hourly resolution, by aggregating the demand from 3 households and pairing with 15 features that include solar PV output, meteorological covariates, lag-based autoregressive terms as well as cyclic hour encodings. Your framework is compared to GRU, vanilla RNN as a benchmark (naive), naive persistence and 24-hour moving average. Key contributions:(i) A unified multi-source feature pipeline for high precision, community-scale off-grid load forecasting;(ii) Systematic LSTM/GRU/RNN comparison on real microgrid data with quantified accuracy gains over statistical baselines; (iii) MAPE reduction of 30.6% over naive persistence (20.94% vs.. 30.16%) and(iv) Feature correlation analysis identifying solar output (r = 0.737), and the lag of one-hour before (r = 0.681), as the dominant predictors.

1.1 Deep Learning for Short-Term Load Forecasting

Stacked Recurrent Models have outperformed ARIMA and exponential smoothing for each demand granularity, meaning the classical statistics methods has since been substantially displaced in favour of deep learning for short-term load forecasting [12]. Against residential smart grid data, comparative evaluations indicate that stacked recurrent architectures achieve a superior accuracy per parameter if the look-back window captures at least one full daily cycle [13]. These gains have recently been extended to multi-horizon forecasting [14] with hybrid CNN-LSTM and attention-augmented models, but hyperparameter sensitivity proves limiting to small off-grid deployments. The block of most benchmark datasets comes from

large utility-scale metropolitan grids, where high data volumes smooth away individual-appliance stochasticity — a condition not satisfied in island microgrids, where household variability is unattenuated and grid import is structurally unavailable [15].

1.2 Recurrent Architectures and Feature Engineering

Stacked LSTMs have shown accurate demand prediction on high frequency residential demand data [15], with gating mechanisms allowing the model to align solar-coupled consumption peaks in off-grid communities [16]. GRU networks provide a competitive accuracy and lower computational cost, as the RMSE gaps are all less than 2% for one-hour-ahead residential forecasting [17]. Cyclic sine/cosine encodings of the hour index [18] mitigate artificial discontinuity at midnight and multi-step lag features {tn−1, tn−2, tn−3, tn−24} are consistently identified as dominant predictors of next-hour demand [19]. Incorporating solar irradiance, temperature, and humidity with lag features achieves substantial reductions in MAPE over demand-only baselines [20]. Most studies instead use meteorological data from remote APIs instead of collocated sensors, introducing the potential for temporal misalignment that does not occur with the Mafate dataset.

1.3 Community-Scale and Off-Grid Microgrid Forecasting

Although both use cases benefit from FSP (Forecast-aware service provisioning), residential load aggregation at community scale yields substantially larger forecast uncertainty than predictions for single-household appliances, due to diversity in occupancy patterns and heterogeneous appliance cycles between households [21]. Data from widespread metering infrastructure can be used for household load forecasting (HLF) without compromising privacy, as shown in several federated and transfer learning approaches [22]. The literature around forecasting clearly states, that Island and off-grid deployments are still explicitly underserved, with AI-based demand forecasting on real off-grid residential data clearly identified as a big gap [15]. Previous studies do not combine real household demand aggregation, spatial co-location of meteorological and solar features, or a systematic architecture comparison across real microgrid scenarios, which is addressed in this paper using an off-grid island microgrid as our case study setting.

II. METHODOLOGY

2.1 Problem Formulation

The forecasting task is one-step ahead: it may be framed as a supervised sequence regression problem, where given 15-dimensional feature vectors for a 24-hour lookback window, the goal is to predict aggregate community load at some point in the future (one hour) L_t (kW):LoadIn: the aggregated active load on the primary bus of three single-family houses. This mapping from input to scalar prediction is implemented using a stacked two-layer LSTM

2.2 Dataset Preprocessing

DatasetThe Mafate offgrid microgrid dataset [11] consists of minute-resolution household electricity measurements from three Reunion Island households (May–December, 2020). The power (W) for the main-bus per household was summed to generate the total load of an entire community, resampled from one-minute to hourly resolution by mean aggregation within each 60-minute window, and converted from watts to kilowatts. The meteorological streams (irradiance, temperature, humidity) and solar plant output are homogenously resampled to the same grain at 1 hour frequency in inner-joined with the demand index resulting in a data set of 5,460 aligned hourly observations between May 2 – December 28, 2020.

2.3 Feature Engineering

It is a 15D vector made of the four groups of features. Meteorological Features: consisting of global horizontal irradiance G_hi (W/m²), ambient temperature T_a (°C) and relative humidity R_h (%) from the collocated station. Solar Generation : PV output power P_solar (kW) from microgrid plant sensor This is where cyclic time features come in; we use sine and cosine projections to encode the hour index h_t, (h_t ∈ {0,...,23}), as: h_t^sin = sin(2\ π h_t/24), h_t^cos = cos(2\ π h_t/24) so we remove that artificial discontinuity at midnight.These are complemented by day-of-week indicator, a weekend flag, and calendar month entries. Autoregressive lag features are used to model low-frequency trends as well as short-range autocorrelation and the intraday periodicity: observed load at lags 1, 2, 3, and 24 hours (short-range auto regression) plus a one-step-shifted (i.e., day before for ground truth) rolling mean (low-frequency trend). All 15 features are min-max normalized with training-set statistics only (no leakage!), while the target load is scaled using a stand-alone scaler and inverse-transformed before running metrics.

2.4 Sequence Construction and Data Split

Sliding-window input-output pairs of length W = 24 are constructed from the 5,460 normalized observations, yielding 5,436 sequences. These are partitioned chronologically: 3,806 training (70%), 815 validation (15%), and 815 test (15%) sequences.

2.6 Model Architecture

The stacked two-layer LSTM processes input tensors of shape (64, 24, 15) — batch × sequence × features. The first LSTM layer (hidden size 128) returns the full sequence to the second layer (hidden size 128, final hidden state only). A dropout layer (p = 0.2) follows, then a fully connected layer (128 → 64, ReLU) and a linear output layer (64 → 1) producing the scalar one-step-ahead prediction. Two deep learning baselines share the identical stacked structure: the GRU baseline replaces LSTM cells with two-gate GRU cells; the vanilla RNN baseline applies simple tanh recurrence. Two statistical baselines complete the comparison: naive persistence ($\hat{L}_{t+1} = L_t$) and a 24-hour rolling mean.

2.7 Training Protocol

All neural models use Adam (η = 0.001) with MSE loss, gradient clipping at norm 1.0, a reduce-on-plateau scheduler (halve η after 5 stagnant epochs), and early stopping (patience = 12, restore best validation weights). Maximum epochs: 80. Table I summarizes all hyperparameters.

Table I: Model Hyperparameters

Parameter	Value
Look-back window W	24 hours
Feature dimension F	15

LSTM/GRU/RNN hidden size H	128
Number of recurrent layers	2
Dropout rate p	0.2
Batch size B	64
Initial learning rate η	0.001
Max training epochs	80
Early stopping patience	12 epochs
LR scheduler patience	5 epochs
LR reduction factor	0.5
Random seed	42
Gradient clip norm	1.0

III. System Modeling and Formulation

3.1 Load Aggregation and Temporal Resampling

All neural models minimize the mean squared error (MSE) loss over mini-batches of B = 64 sequences:

$$\mathcal{L}(\theta) = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{y}_i - \tilde{L}_{i+W}\right)^2 \qquad (1)$$

where y^_i is the normalized prediction for sequence i and L~_(i+W) is the normalized ground-truth load. Parameters θ are updated by the Adam optimizer with learning rate η=0.001, first-moment decay β_1=0.9, and second-moment decay β_2=0.999 [20]. Gradient norms are clipped to a maximum of 1.0 to prevent parameter instability:

$$\mathbf{g}^{\text{clip}} = \mathbf{g} \cdot \min\left(1, \frac{1.0}{\| \mathbf{g} \|_2}\right) \qquad (2)$$

$$\hat{y}_i = \mathbf{w}_2^{\top}\text{ReLU}(\mathbf{W}_1\mathbf{h}_W + \mathbf{b}_1) + b_2 \qquad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{64\times128}$, $\mathbf{w}_2 \in \mathbb{R}^{64}$, and $b_1 \in \mathbb{R}^{64}$, $b_2 \in \mathbb{R}$ are the dense layer parameters.

IV. Results and Discussions

We use the processed Mafate dataset [11] provides 5,460 hourly observations (May 2 – December 28, 2020), consisting of aggregate community load ranging from a minimum power of 0.034 kW to maximum of 1.021 kW and then an average value equal to 0.193 kWdisworking hours. The load profile is characterized by a pronounced bimodal intraday pattern: demand troughs at 05:00 (0.128 kW) and peaks 19:00 (0.261 kW), whilst a 73.3% morning ramp from the fixed scheduling of solar-coupled cooking activity drives mean demand up from 07:00 to 13:00. Fig. The structure is shown in 2a with shading at ±1 standard deviation. We find that solar plant output is the strongest cross-feature predictor (r = 0.737) across all measurements: this captures behavioral coupling between PV availability and appliance activity through feature correlation analysis. Strong autoregressive structure is confirmed by the 1-hour load lag (r = 0.681) and the 24-hour lag (r = 0.606). Results showed seasonal signal in calendar month (r = 0·489) and ambient temperature (r = 0·426). Correlation between day-of-week (r = −0.007) and weekend flag (r = −0.020) is weak — as a permanently off-grid settlement it does not follow conventional weekly cycles in terms of occupancy patterns.

4.1 Comparative Forecasting Performance:

Accuracy of all 5 models on 815 test sequences is reported in Table II. Both statistical baselines are decisively outperformed by all three deep learning architectures. The LSTM produces RMSE = 0.0856 kW and MAPE = 20.94% — deviations of +20.1%, −30.6% over naive persistence (RMSE = 0.1072 kW, MAPE = 30.16%). GRU has a RMSE = 0.0845 kW and MAPE = 19.96% while the vanilla RNN is the best for RMSE (0.0826 kW) and MAE (0.0502 kW), out of all models.

Table II: Forecasting Accuracy — All Models (Test Set, N=815)

Model	RMSE (kW)	MAE (kW)	MAPE (%)	R^2
LSTM	0.0856	0.0527	20.94	0.3645
GRU	0.0845	0.0511	19.96	0.3807
Vanilla RNN	0.0826	0.0502	20.38	0.4093
Naive Persistence	0.1072	0.0695	30.16	0.0046
Moving Average (24 h)	0.1047	0.0755	33.26	0.0550

With the RMSE = 0.1047 kW, MAPE = 33.26%, the final model of a naive persistence mean with a duration of two days (24 hour) improves marginally on naively smoothing peak demand in term of RMSE but worsens it for MAPE, verifying that is over-attenuating sharp spikes in demand. The two R^2 values of 0.0046 and 0.0550 for the statistical baselines demonstrate that neither captures non-random structure in load variance, while deep learning models explain 36.5%–40.9% of test variance.

The 3.5% spread across architectures comes into play when we look at the RMSE ordering: vanilla RNN (0.0826) < GRU (0.0845) < LSTM (0.0856). The simpler architecture generalizes better when evaluated on 3,806 training sequences from a single community: the vanilla RNN has lower parameter count than its high-dimensional counterpart, making it less prone to over-fitting this relatively small dataset. The gating advantage of the LSTM is strongest on multi-day horizons; it matters less for the 24-hour-back test used by this study. As can be observed through all three architectures, MAPE remains significantly below both the statistical baselines (best DL: 19.96% vs best statistical:30.16%), confirming that our meta-feature pipeline drives accuracy improvements irrespective of architecture selection. Fig. The complete comparison of each metric for all five models can be seen on 2(b). Comparative bar chart of RMSE, MAE, and MAPE across all five models (F0igure 1, right panel) intuitively displays the accuracy gap between AI-powered versus statistical approaches. The deep learning cluster fills the left side of the chart with all error bars lower than those for the statistical baselines, which appear significantly higher on all three metrics.

FIGURE 2: (a) Hourly Load Profile — Mean aggregate community load (kW) with ±1 standard deviation shading across 24 hours of the day, showing bimodal structure with evening peak at 19:00 (0.261 kW) and predawn trough at 05:00 (0.128 kW). (b) Model Comparison — Bar chart of

RMSE, MAE, and MAPE/100 for LSTM, GRU, Vanilla RNN, Naive Persistence, and Moving Average (24 h) on 815 test sequences. MAPE values annotated above respective bars.

4.2 Temporal Load Demand Patterns and Feature Relevance

Table III shows hourly load statistics for some selected hours. The maximum stress hour (19:00; 0.261 kW) and minimum hour of the day (0.128 kW at 05:00) define a 50.9% reduction from daily maximum, after which battery discharge is typically minimized overnight in off-grid configuration The standard deviation is at maximum levels between 08:00–09:00; it reaches 0.113–0.156 kW in this time horizon and the single-session maximum of 1.021 kW occurred —that is, the window where solar irradiance onset triggers stochastic cooking activity. The LSTM accurately predicts this morning ramp from the profile of the prior day, thanks to its dependence on 24-hour lag (r = 0.606) and rolling mean (r = 0.577), which allows it to see into tomorrow's forecasted values, providing an advantage over statistical methods in many high-variance hours.

Table III: Hourly Load Statistics — Aggregate Community Demand (Selected Hours)

Hour	Mean (kW)	Std (kW)	Min (kW)	Max (kW)	Peak Reduction (%)
00	0.1519	0.0555	0.0523	0.3683	41.8
05	0.1281	0.0460	0.0406	0.2763	50.9
08	0.1683	0.1126	0.0394	1.0214	35.5
12	0.2425	0.0994	0.0436	0.4949	7.1
13	0.2489	0.1068	0.0517	0.7253	4.6
19	0.2610	0.0908	0.0728	0.4803	0.0
21	0.2092	0.0787	0.0531	0.5206	19.8
23	0.1659	0.0614	0.0499	0.3682	36.4

A large standard deviation at 08:00-09:00 (0.113 kW and 0.156 kW respectively), as well an extreme maximum of 1.021 kW recorded at this time of day, illustrate the window period with the highest forecast uncertainty level — Table 1, Figs. These hours coincide with the onset of solar irradiance () and the start of individual cooking activity — a behavioural event much more stochastically timed and emitted than, say, cellular respiration. This window captures the solar–demand coupling supported by feature correlation () more than any other time frame. In addition to this, the character of the LSTM encodes in the pin 24h lag term (,) as well as for rolling mean (,) that enable it to anticipate such morning ramp from yesterday profile and thereby outperform purely statistical methods at high-variance hours.

4.3 Training Convergence and Computational Performance

The training and validation MSE loss curves for the three separate deep learning architectures are shown in Figure 2. Early stopping and ReduceLROnPlateau scheduling allow all three models to converge stably. The best validation loss for the LSTM is 0.003796 after 58 epochs; the GRU, converging fastest with its simpler gate structure, achieves a loss of 0.004006 at epoch 35; and vanilla RNNs, in contrast each train through for 41 epochs (0.003885). All three exhibit a steep drop in loss during the first five epochs, followed by a plateau phase where the learning rate gets halved by the scheduler. This results in the LSTM having the lowest validation loss but ranks third for test RMSE — a swing that can be traced to seasonal distribution shift from the validation period (October 2020) to the test period (November–December 2020), where daylight hours and solar irradiance vary. The test generalization of a vanilla RNN is worse due to it having fewer parameters to overfit the validation distribution. Timing for training on CPU: LSTM 8 min, GRU 4 min, RNN ~5 min (batch size 64) Fig. The entire loss trajectories are demonstrated in Fig. 3

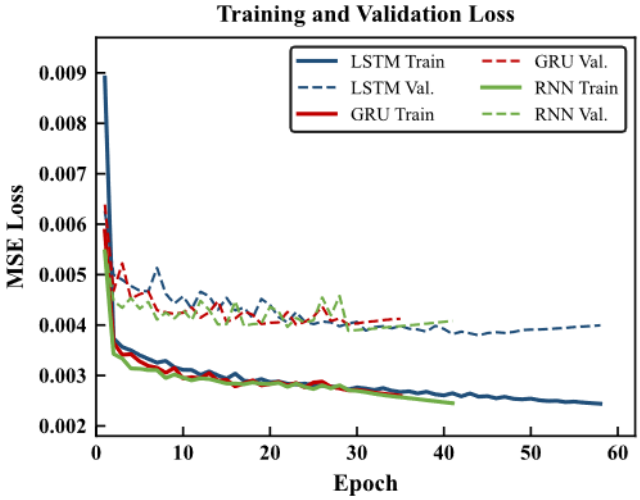

FIGURE 3: Training and Validation MSE Loss Curves — Training (solid) and validation (dashed) loss trajectories for LSTM (blue, 58 epochs), GRU (red, 35 epochs), and vanilla RNN (green, 41 epochs). All models trained with early stopping (patience=12), Adam optimizer, and ReduceLROnPlateau scheduling. Best validation losses: LSTM=0.003796, GRU=0.004006, RNN=0.003885.

CONCLUSION

A stacked two-layer LSTM load forecasting framework was evaluated on the real-life Mafate off-grid residential microgrid dataset in this paper. With a 24-hour look-back, an input pipeline was constructed using autoregressive load lags, meteorological covariates, solar generation and cyclic hour encodings; the neural net was trained with Adam optimization, early stopping and adaptive learning rate scheduling across 15 features. On 815 test sequences LSTM RMSE = 0.0856 kW, MAE = 0.0527 kW and MAPE=20.94% which are reductions of RMSE: -20.1%, MAPE:-30.6% against naive persistence. A RMSE span of 0.0826–0.0856 kW was obtained by all three deep learning architectures, outperforming both statistical baselines (0.1047–0.1072 kW for persistence and moving average). The strongest demand predictors were solar plant output (r = 0.737) and 1-hour lag (r = 0.681); ambient temperature (r = 0.426) and calendar month (r = 0.489) provided complimentary seasonal signal to these variables.

Future work include (i) extending to multi-step forecast horizons (6, 12, 24 hours)—for intraday dispatch planning;

(ii) integrating attention mechanisms and transformer architectures on small microgrid datasets; (iii) incorporating battery state-of-charge as an input feature for forecast-aware charge–discharge optimization; (iv) leveraging probabilistic frameworks — quantile regression and Bayesian LSTM — for prediction intervals under demand-response uncertainty; and finally, (v) validating the feature pipeline on additional off-grid island deployments to evaluate generalization.

REFERENCES

[1] Biswal, Biswajit, Subhasish Deb, Subir Datta, Taha Selim Ustun, and Umit Cali. "Review on smart grid load forecasting for smart energy management using machine learning and deep learning techniques." Energy Reports 12 (2024): 3654-3670.

[2] Yu, Kun. "Adaptive Bi-directional LSTM short-term load forecasting with improved attention mechanisms." Energies 17, no. 15 (2024): 3709.

[3] Gassar, Abdo Abdullah Ahmed. "Short-term energy forecasting to improve the estimation of demand response baselines in residential neighborhoods: Deep learning vs. machine learning." Buildings 14, no. 7 (2024): 2242.

[4] Jrhilifa, Ismael, Hamid Ouadi, Abdelilah Jilbab, Nada Mounir, and Abdellah Ouaguid. "A VMD-deep learning approach for individual load monitoring and forecasting for residential buildings energy management." e-Prime-Advances in Electrical Engineering, Electronics and Energy 8 (2024): 100624.

[5] Lu, Shuai, and Taotao Bao. "Short-term electricity load forecasting based on NeuralProphet and CNN-LSTM." IEEE access 12 (2024): 76870-76879.

[6] Singh, Gurjot, and Jatin Bedi. "A federated and transfer learning based approach for households load forecasting." Knowledge-Based Systems 299 (2024): 111967.

[7] Deka, Priyam, Moisés Cordeiro-Costas, Raquel Pérez-Orozco, Michał Chabiński, and Andrzej Szlęk. "Novel NSGA-II optimized LSTM and GRU models for short-term forecasting of residential heating load." Energy and Buildings 344 (2025): 115999.

[8] Zhou, Liangcai, Yi Zhou, Linlin Liu, and Xiaoying Zhao. "A comparative study of different deep learning methods for time-series probabilistic residential load power forecasting." Frontiers in Energy Research 12 (2024): 1490152.

[9] Iqbal, Muhammad Sajid, and Muhammad Adnan. "Edge computing and transfer learning-based short-term load forecasting for residential and commercial buildings." Energy and Buildings 329 (2025): 115273.

[10] Sun, Yuting, Zhipeng Chang, Jianan Yu, and Zongxiang Chen. "Blockchain-Enabled Microgrid IoT with Accurate Predictions of Renewable Energy and Electricity Load Using LevySSA-LSTM-GRU." Sustainability 18, no. 3 (2026): 1653.

[11] Calogine, Didier, Johann Francou, Cedric Abbezzot, and Tovondahiniriko Fanjirindratovo. "Data in experimental stand-alone microgrid: Solar production, domestic loads, battery storage and meteorological series." Data in Brief 51 (2023): 109643.

[12] Aravanis, Theofanis, and Andreas Kanavos. "Comparative Evaluation of Deep Learning Architectures for Electricity Demand Forecasting." Mathematics 14, no. 5 (2026): 827.

[13] Baek, Minsoo, and Youngguk Seo. "Hybrid forecasting of university electricity demand using time series and deep learning." Energy and Buildings (2025): 116400.