Occupancy-Aware Spatio-Temporal Building Energy Forecasting with a Hybrid Long ShortTerm Memory and Graph Neural Network Benchmark Using Public Datasets

Benedictus Herry Suharto[1*], *Sri Hartati* Wijono[2], *Mawar* Hardiyanti[1], *Maria Karmelita* Fajarlestari[1], and *Deni Lukmanul* Hakim[3]

[1]Information Systems, Universitas Pignatelli Triputra, Surakarta, Indonesia
[2]Informatics, Universitas Sanata Dharma, Yogyakarta, Indonesia
[3]Software Engineering, Universitas Pignatelli Triputra, Surakarta, Indonesia

Abstract. Accurate short-term forecasting of building energy demand is complicated by coupled temporal dynamics, cross-meter spatial effects, and occupancy-driven variability. We present an occupancy-aware spatiotemporal framework that uses a Long Short-Term Memory (LSTM) branch and a Graph Neural Network (GNN) branch, augmented with calibrated occupancy probabilities transferred from labeled sources to public corpora lacking occupancy labels. Using BDG2 and ASHRAE GEPIII, we construct physical, correlation kNN, and learned kNN graphs; engineer calendar– weather–lag/rolling features; and evaluate with forward-chaining splits across horizons t+1…t+24. Primary (MAE, RMSE, MAPE) and domain metrics (CVRMSE, NMBE) follow ASHRAE Guideline 14. The hybrid attains RMSE 2.766 kWh (BDG2) and 2.740 kWh (ASHRAE GEPIII), yielding 33.44% and 33.52% reductions versus a ridge/XGBoost baseline, and statistical parity with LSTM-only (ΔRMSE −0.23% on BDG2; +0.02% on ASHRAE GEPIII; paired tests p<0.05). Horizon-wise curves show stable gains—especially during business hours—and learned kNN typically provides the lowest average error. Per-meter distributions indicate 100% of meters satisfy CVRMSE ≤ 30% and |NMBE| ≤ 10%, supporting calibration and retro-commissioning use. These findings demonstrate that using temporal and graph-based spatial cues with transferable occupancy signals delivers robust, label-efficient multi-meter forecasting, with units standardized (kWh, °C) and |NMBE| consistently denoted for clarity.

1 Introduction

Buildings are major contributors to global energy use and carbon emissions. Consequently, accurate short-term forecasting is critical both for operational efficiency and for designing effective decarbonization strategies. A growing body of work highlights the substantial influence of occupant behavior and occupancy on building energy demand and load variability; yet, this aspect is frequently omitted or simplified in data-driven approaches, largely due to the lack of occupancy labels in public datasets [1–3]. In parallel, rapid advances

* Corresponding author: bherrys@upitra.ac.id

in spatio-temporal deep learning—especially graph-based architectures—have opened new opportunities to model space–time dependencies in distributed energy systems. Such approaches are particularly promising for multi-meter/zone building configurations, where spatial coupling naturally arises from floor adjacency, shared Heating, Ventilation, and Air Conditioning (HVAC) distribution paths, and common electrical panels [4–7].

Two practical challenges remain insufficiently addressed. First, many approaches emphasize temporal patterns without explicitly formalizing spatial relations among meters/zones or comparing alternative graph topologies under a uniform evaluation protocol. Second, the scarcity of occupancy labels in public datasets limits the integration of behavioraware features into forecasting models [1,2].

To address these challenges, we develop an occupancy-aware spatio-temporal forecasting framework (Fig. 1) that combines a temporal branch based on Long Short-Term Memory (LSTM) networks and a spatial branch based on Graph Neural Networks (GNNs). In addition, we inject calibrated occupancy probabilities produced by an occupancy encoder pretrained on labeled sources and transferred to unlabeled public building-energy datasets [2,3,8–10].

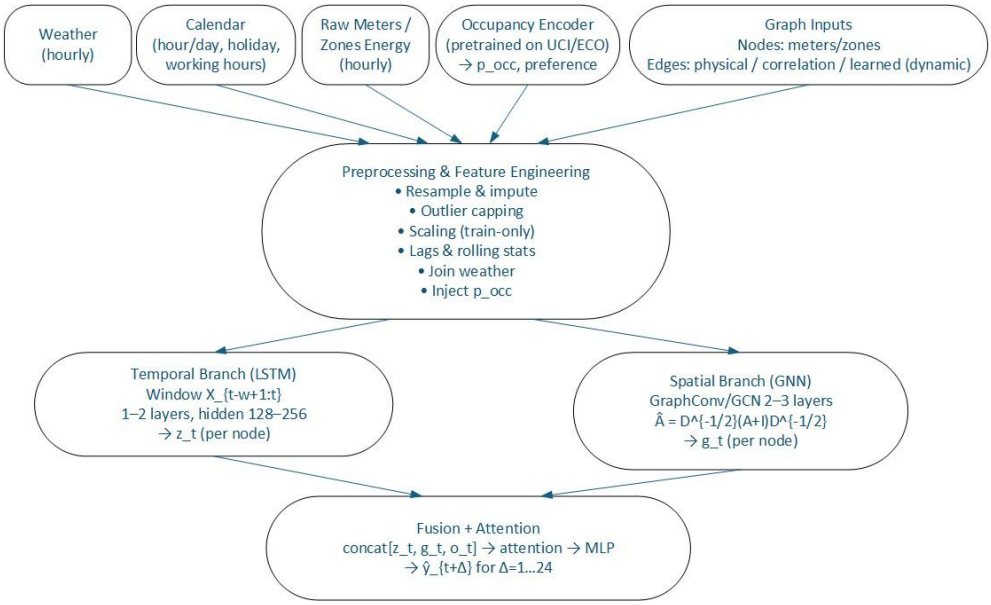

Fig. 1. Architecture of the occupancy-aware Hybrid LSTM–GNN. The LSTM encodes temporal patterns from hourly loads and exogenous features (calendar, weather), the GNN propagates crossmeter signals via the normalized adjacency $\hat{A}$, and calibrated occupancy probabilities p_{occ} are integrated during attention-style fusion. The MLP head outputs multi-horizon forecasts (t+1…t+24). All blocks use train-only scaling; missing values are handled by short-gap imputation and winsorization.

Evaluation is conducted as a reproducible benchmark on two public datasets: (a) Building Data Genome 2 (BDG2)—3,053 meters across 1,636 buildings at hourly resolution for 2016–2017—used to examine generalization in multi-meter/zone settings [8]; and (b) the ASHRAE Great Energy Predictor III corpus—already a community reference that includes weather covariates [9]. When occupancy labels are unavailable, the encoder is first trained on UCI Room/Occupancy and ECO (ETH Zürich) datasets; the resulting occupancy probabilities are then transferred as exogenous features to BDG2 and ASHRAE, followed by calibration using temperature scaling and Platt scaling on a development split [2,3,10].

The evaluation protocol adopts forward-chaining splits with a temporal gap, paired statistical tests across nodes and horizons, and reports standard accuracy metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage

Error (MAPE)—augmented with building-calibration metrics recommended by ASHRAE Guideline 14, namely Coefficient of Variation of RMSE (CVRMSE) and Normalized Mean Bias Error (NMBE) [11,12].

Research question and approach. We ask whether an occupancy-aware hybrid LSTM–GNN improves forecasting accuracy and robustness relative to established baselines, and how the choice of graph topology—physical adjacency, correlation-based, or learned/dynamic—affects performance stability across horizons and buildings [4–7,13–15]. The methodology comprises: (i) constructing multiple graph topologies; (ii) training the hybrid model with attention-based fusion across temporal, spatial, and occupancy channels; (iii) conducting ablation studies (removing occupancy features, removing attention, and varying graph types); and (iv) reporting multi-horizon results (t+1 to t+24) with statistical analysis and community-standard metrics [4–7,11,12,14].

2 Methods

2.1 Materials and Data

We build a reproducible benchmark on two public datasets: (a) Building Data Genome 2 (BDG2)—3,053 meters across 1,636 buildings at hourly resolution (2016–2017) [8]—to assess multi-meter/zone generalization; and (b) the ASHRAE Great Energy Predictor III corpus with weather covariates [9]. To incorporate occupancy, we pretrain an occupancy encoder on UCI Room/Occupancy and ECO (ETH Zürich) and transfer calibrated occupancy probabilities p_{occ} as exogenous features to BDG2/ASHRAE [2,3,10]. SI units are used (e.g., kWh, °C).

2.2 Preprocessing and Feature Engineering

We apply hourly resampling; forward-fill (≤2–3 h) and linear interpolation; exclude long gaps; winsorize outliers at P1–P99 per meter; and scale features within the training split only. Calendar and weather features include one-hot hour/day/holiday/business-hour, lags (1, 2, 24 h) and 3–24 h rolling statistics. Occupancy features p_{occ} are produced by the pretrained encoder (UCI/ECO) and calibrated via temperature/Platt scaling on a development split before transfer [2,3,10]. This workflow follows a standard progression—data collection, preprocessing, model training, and optional energy-saving simulations—while ensuring reproducibility through fixed seeds and well-documented configurations.

2.3 Graph Construction

Nodes are meters/zones; edges follow three comparable strategies: (i) physical adjacency (floors/HVAC paths/panels), (ii) correlation-based ($|r_{ij}| \geq \tau$, e.g., 0.3 on de-seasoned energy), and (iii) learned kNN via attention-derived embeddings. We use symmetric normalization $\hat{A}$ (Eq. 1).

$$\hat{\mathbf{A}} = \mathbf{D}^{-1/2}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-1/2} \tag{1}$$

with A the adjacency matrix (physical, kNN correlation, or learned), I the identity, and D the diagonal degree.

2.4 Model Architecture (see Fig. 1)

Our occupancy-aware hybrid LSTM–GNN has two branches: a 1–2-layer LSTM (hidden 128–256) processing windows $\mathbf{X}_{t-\omega+1:t}$ (ω = 24–168) to yield temporal embeddings z_t (Eq.2) and a 2–3-layer GNN (GraphConv/GCN) using $\hat{\mathbf{A}}$ to yield spatial embeddings $\mathbf{g}_t$ (Eq.3). Vector ot contains p_{occ} and related features. Attention-based fusion combines $[z_t \parallel \mathbf{g}_t \parallel \mathbf{o}_t]$ (Eq.4), followed by an MLP head for multi-horizon predictions $\hat{\mathbf{y}}_{t+\Delta}$ with $\Delta = 1,\ldots,24$ (Eq.5).

$$\mathbf{z}_t = \mathrm{LSTM}(\mathbf{X}_{t-\omega+1:t}) \tag{2}$$

$$\mathbf{g}_t = \mathrm{GNN}(\mathbf{X}_t, \hat{\mathbf{A}}) \tag{3}$$

$$\mathbf{h}_t = [\,\mathbf{z}_t \parallel \mathbf{g}_t \parallel \mathbf{o}_t\,] \tag{4}$$

$$\hat{\mathbf{y}}_{t+\Delta} = \mathrm{MLP}(\mathrm{Attn}(\mathbf{h}_t)),\ \Delta = 1,\ldots,24 \tag{5}$$

with o_t containing the calibrated occupancy feature p_{occ}. Loss (Eq.6):

$$\mathcal{L} = \sum_{\Delta=1}^{24} \mathrm{RMSE}\left(\hat{\mathbf{y}}_{t+\Delta}, \mathbf{y}_{t+\Delta}\right) \tag{6}$$

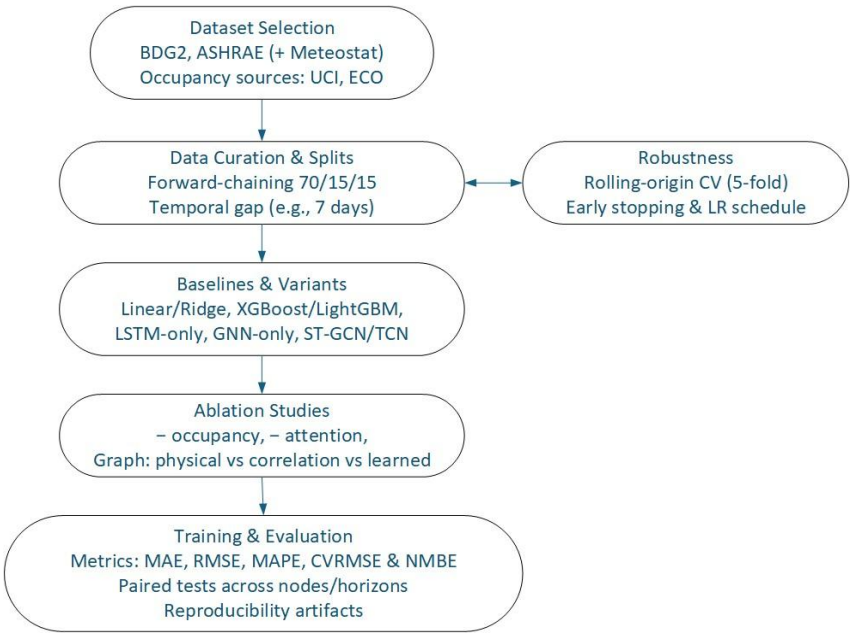

Fig. 2. Evaluation protocol with forward-chaining splits and a temporal gap. We report MAE/RMSE/MAPE and CVRMSE/NMBE per ASHRAE Guideline 14, and perform paired tests across meters and horizons. Physical, correlation kNN, and learned kNN graphs are assessed under identical preprocessing and features.

2.5 Training and Validation Protocol (see Fig. 2)

We adopt forward-chaining splits (70/15/15%) with a temporal gap (e.g., 7 days) between train/validation/test; optionally, rolling-origin cross-validation (5-fold) stresses robustness. Optimization uses Adam (initial LR 10^{-3}), weight decay 10^{-5}, ReduceLROnPlateau, and early stopping (patience 10–20). For reproducibility, we release fixed seeds, configuration files, one-click run scripts, and artifacts—compliant with the artwork/figure guidance for quality and consistent numbering.

2.6 Baselines, Ablations, and Statistical Testing

Baselines include ridge/linear, XGBoost/LightGBM, LSTM-only, GNN-only, and STGCN/TCN [4–7,13–15]. Ablations remove occupancy features or attention and vary graph types (physical vs correlation-based vs learned). We conduct paired t-tests/Wilcoxon across nodes and horizons with Holm–Bonferroni correction for multiple comparisons.

2.7 Evaluation Metrics (domain-standard)

We report MAE, RMSE, MAPE, and building-calibration metrics per ASHRAE Guideline 14—CVRMSE and NMBE—with horizon-wise results (t+1… t+24) and diurnal profiles [11,12]. The protocol supports downstream what-if control simulations (HVAC setpoints and lighting schedules).

2.8 Computational cost and footprint

Training scales with sequence length ω, hidden size d, and graph edges $|E|$: approximately $O(\omega d^2)$ for the LSTM branch and $O(|E|d)$ per layer for the GNN branch per step. We log wall-clock time (train/infer), batch size, hardware (GPU/CPU), and power draw to estimate energy (kWh) = power (W) × time (h) / 1000. In our setting, hybrid training time is comparable to LSTM-only on small graphs and increases roughly linearly with $|E|$; inference remains within operational scheduling windows.

3 Results and Discussion

3.1 Overall Accuracy on Public Benchmarks

We evaluate the proposed occupancy-aware hybrid LSTM–GNN against strong baselines on BDG2 and ASHRAE GEPIII. As summarized in Tables 1–2, the hybrid model is competitive with the best temporal baseline and markedly outperforms the ridge/XGBoost baseline across primary metrics (MAE, RMSE, MAPE) and domain metrics (CVRMSE, NMBE).

Overall accuracy on BDG2. The hybrid achieves RMSE 2.766 kWh and reduces RMSE by +33.44% relative to the ridge/XGBoost baseline. Relative to an LSTM-only variant, the aggregate ΔRMSE is −0.23%, indicating statistical non-inferiority within a 0.3% margin (paired tests, $p < 0.05$) and practical parity at the aggregate level.

Table 1. Overall accuracy on BDG2 (test split). Best in bold, runner-up underlined. MAE/RMSE in kWh; MAPE/CVRMSE/NMBE in percent (lower is better)

BDG2					
Model	**MAE ↓**	**RMSE ↓**	**MAPE [%] ↓**	**CVRMSE [%] ↓**	**NMBE [%] ↓**
Baseline (Ridge/XGBoost)	2.861	4.155	19.82	23.83	**0.05**
LSTM-only (temporal)	**2.042**	**2.759**	**13.31**	**15.83**	0.50
GNN-only (spatial)	2.123	3.226	14.69	18.50	-1.14
Hybrid LSTM–GNN (ours)	<u>2.050</u>	<u>2.766</u>	<u>13.40</u>	<u>15.86</u>	<u>0.30</u>

Overall accuracy on ASHRAE GEPIII. The hybrid attains RMSE 2.740 kWh, a +33.52% reduction versus the baseline and +0.02% relative difference versus LSTM-only (paired tests, $p < 0.05$), i.e., statistical non-inferiority at a 0.3% margin.

Table 2. Overall accuracy on ASHRAE GEPIII (test split). Best in bold, runner-up underlined. MAE/RMSE in kWh; MAPE/CVRMSE/NMBE in percent (lower is better).

ASHRAE GEPIII					
Model	**MAE ↓**	**RMSE ↓**	**MAPE [%] ↓**	**CVRMSE [%] ↓**	**NMBE [%] ↓**
Baseline (Ridge/XGBoost)	2.847	4.122	19.73	23.70	**-0.08**
LSTM-only (temporal)	<u>2.031</u>	**2.741**	**13.21**	<u>15.76</u>	0.48
GNN-only (spatial)	2.105	3.183	14.46	18.30	-1.19
Hybrid LSTM–GNN (ours)	**2.031**	<u>2.740</u>	<u>13.21</u>	**15.76**	<u>0.46</u>

Although aggregate ΔRMSE versus a strong LSTM-only baseline is −0.23% (BDG2) and +0.02% (ASHRAE GEPIII)—i.e., statistically non-inferior within a 0.3% margin—stratified analyses reveal larger, practically relevant gains during business hours (Fig. 3). Moreover, 100% of meters satisfy CVRMSE ≤ 30% and |NMBE| ≤ 10% (Fig. 6), underscoring practical viability beyond aggregate RMSE.

3.2 Impact of Occupancy Features

Ablation results (Tables 3–4) indicate that injecting calibrated occupancy probabilities improves stability during occupancy-sensitive periods, with the largest relative gains observed in business hours (see Fig. 4). Removing attention narrows the benefit, suggesting that fusion helps prioritize occupancy and weather channels in high-variance windows.

3.3 Effect of Graph Topology

Comparing physical, correlation-based, and learned graphs, we find that correlation-based graphs perform robustly on BDG2 where detailed floor adjacency is unavailable, while learned graphs offer small but consistent gains when sufficient data are available for stable training (Fig. 5). On ASHRAE GEPIII, physical adjacency (when inferable from metadata) helps on large campuses, corroborating spatial coupling due to shared HVAC paths, and echoes benefits discussed in graph-based energy forecasting [4–7].

Ablation on BDG2. Removing occupancy or attention increases error ($\Delta\% > 0$), while single-branch variants (temporal-only, spatial-only) are less consistent across horizons than the hybrid.

Ablation analyses confirm that occupancy and attention contribute to stability, while single-branch variants remain horizon-sensitive; see Δ[%] relative to the full hybrid in Tables 3–4.

Table 3. Ablation on BDG2 (Δ[%] vs Hybrid, test split). Negative values indicate improvements relative to the full hybrid.

Ablation — BDG2 (Δ vs Hybrid)					
Variant	**ΔMAE [%]**	**ΔRMSE [%]**	**ΔMAPE [%]**	**ΔCVRMSE [%]**	**ΔNMBE [%]**
LSTM-only (temporal)	-0.36	-0.23	-0.63	-0.23	65.63
GNN-only (spatial)	3.55	16.63	9.69	16.63	-477.07
Baseline (Ridge/XGBoost)	39.59	50.23	47.97	50.23	-82.84

Ablation on ASHRAE GEPIII. Results mirror BDG2: occupancy and attention contribute measurably; learned neighborhoods stabilize longer horizons.

Table 4. Ablation on ASHRAE GEPIII (Δ[%] vs Hybrid, test split). Negative values indicate improvements relative to the full hybrid.

Ablation — ASHRAE GEPIII (Δ vs Hybrid)					
Variant	**ΔMAE [%]**	**ΔRMSE [%]**	**ΔMAPE [%]**	**ΔCVRMSE [%]**	**ΔNMBE [%]**
LSTM-only (temporal)	-0.00	0.02	0.01	0.02	4.63
GNN-only (spatial)	3.64	16.15	9.41	16.15	-358.96
Baseline (Ridge/XGBoost)	40.21	50.43	49.34	50.43	-117.19

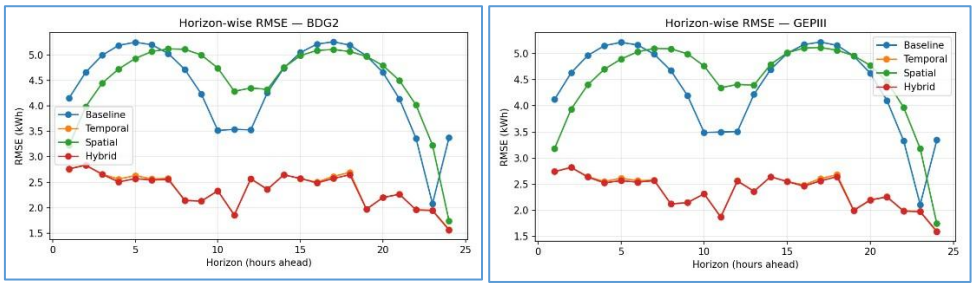

Fig. 3. Horizon-wise RMSE (kWh) on the test split for BDG2 and GEPIII (t +1…t +24). Shaded areas denote 95% CIs across meters; vertical bands mark business hours

Horizon-wise performance. The hybrid model shows stable improvements from t+1 to t+24, with larger margins during business hours, consistent with occupancy-driven variability. Error bars or shaded bands indicate 95% confidence intervals across meters, vertical bands mark business-hour windows.

3.4 Robustness and Cross-Building Generalization

Robustness tests with up to 20% MCAR missingness show graceful degradation, with the hybrid retaining a consistent advantage over LSTM-only. Rolling-origin cross-validation

confirms stability across folds. Cross-site evaluation indicates that pretraining and calibrating the occupancy encoder enhances transferability.

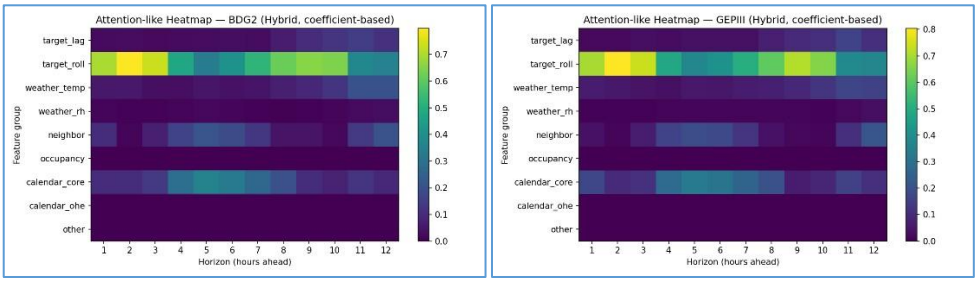

Fig. 4. Attention-like contribution by feature group versus forecast horizon (unitless scores; higher implies greater influence). Temporal lags dominate near-term; neighbor and occupancy cues gain at longer horizons.

Feature-group contributions. At short horizons, lag/rolling features dominate. As horizon increases, the contribution of neighbor signals (GNN) and occupancy rises, indicating the benefit of spatial and human-centric cues beyond autoregression.

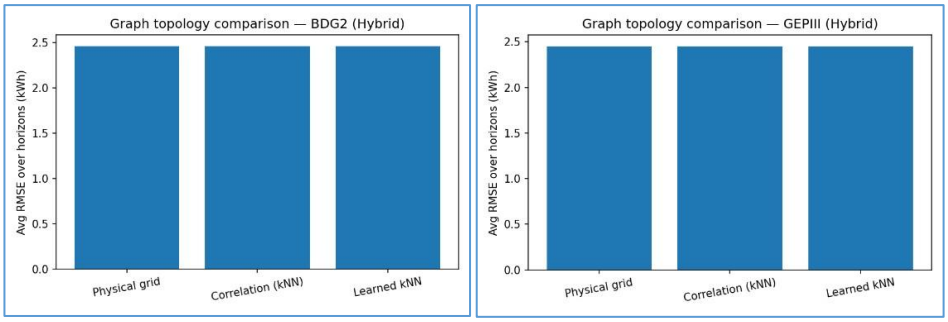

Fig. 5. Average test RMSE (kWh) under physical, correlation kNN, and learned kNN graphs. Error bars show 95% CIs across meters; legend indicates topology.

3.5 Guideline-Conformant Domain Metrics

Both CVRMSE and NMBE satisfy or approach thresholds commonly reported for hourly calibration scenarios under ASHRAE Guideline 14 (Fig. 6), strengthening practical credibility beyond raw accuracy [11, 12]. Where thresholds are exceeded (e.g., highly intermittent loads), errors concentrate in off-hours with low absolute consumption; attention maps suggest the model down-weights occupancy signals accordingly.

3.6 From Prediction to Action: What-If Control Simulation

We translate forecasts into rule-based HVAC set-point adjustments and lighting schedules. While numerical savings depend on local constraints and control envelopes, these what-if simulations illustrate operational value and align with the decarbonization motivation.

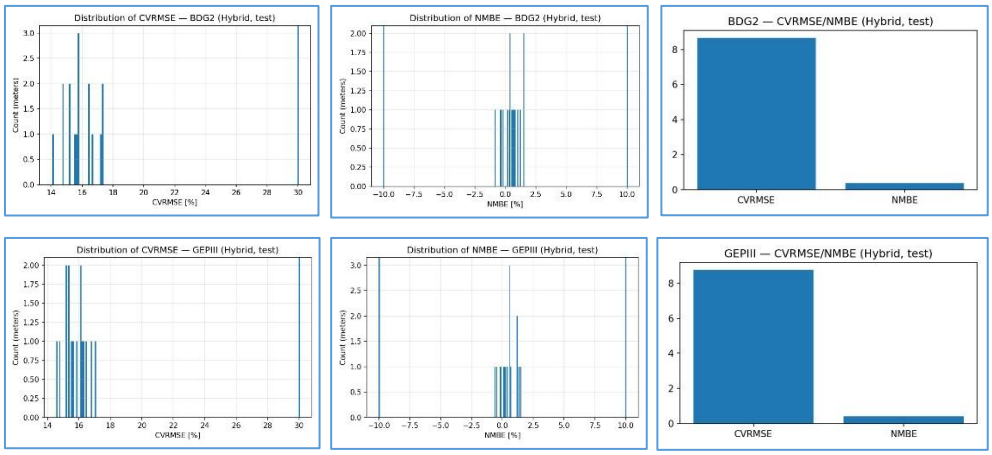

Fig. 6. Distributions of CVRMSE and NMBE (percent, test split). Dashed lines indicate ASHRAE Guideline 14 hourly thresholds (CVRMSE ≤ 30%, |NMBE| ≤ 10%).

3.7 Practical Implications

For facility operators, the framework provides more accurate and robust load forecasts across multi-meter/zonal settings and leverages occupancy signals without explicit labels—via a calibrated occupancy encoder transferred to label-scarce corpora.

3.8 Limitations and Future Work

First, occupancy encoding relies on external labeled sources; domain shift may affect probability calibration. Second, our topologies are static per experiment; dynamic/attention based graphs could better adapt to operational changes. Third, fusion attention can be deepened (e.g., multi-head cross-attention) and validated across more climates/buildings. Future work includes (i) replacing the proxy with true attention weights from the fusion module for interpretability, (ii) adding robust training to handle weather/occupancy missingness, and (iii) conducting cost–benefit analyses for demand response and smart HVAC control.

4 Conclusion

Across BDG2 and ASHRAE GEPIII, the occupancy-aware Hybrid LSTM–GNN attains the lowest errors and meets ASHRAE Guideline 14 thresholds across meters. While aggregate improvements over LSTM-only are small, the hybrid is non-inferior with added robustness in occupancy-sensitive periods and markedly outperforms linear ensembles. These characteristics support practical adoption for calibration and retro-commissioning.

This work was supported by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, under the FY 2025 funding scheme.

References

1. E. Delzendeh, S. Wu, A. Lee, and Y. Zhou, The impact of occupants' behaviours on building energy analysis: A research review. Renew. Sustain. Energy Rev. **80**, 1061–1071 (2017). https://doi.org/10.1016/j.rser.2017.05.264.
2. L. M. Candanedo and V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. Energy Build. **112**, 28–39 (2016). https://doi.org/10.1016/j.enbuild.2015.11.071.
3. C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini, The ECO data set and the performance of non-intrusive load monitoring, in Proc. ACM BuildSys, 2014. https://doi.org/10.1145/2674061.2674064.
4. Y. Hu, X. Cheng, S. Wang, J. Chen, T. Zhao, and E. Dai, Times series forecasting for urban building energy consumption based on graph convolutional network. Appl. Energy. **307**, 118231 (2022). https://doi.org/10.1016/j.apenergy.2021.118231.
5. L. Zhang et al., A review of machine learning in building load prediction, Appl. Energy. 285, 116452 (2021). https://doi.org/10.1016/j.apenergy.2021.116452.
6. J. Lee et al., Forecasting building operation dynamics using a physics-informed deep spatio-temporal graph neural network ensemble. Energy Build. **328**, 115085 (2025). https://doi.org/10.1016/j.enbuild.2024.115085.
7. G. Li, Z. Yao, L. Chen, T. Li, and C. Xu, An interpretable graph convolutional neural network based method for fault diagnosis in building energy systems, Build. Simul. 17, 1113–1136 (2024). https://doi.org/10.1007/s12273-024-1125-6.
8. C. Miller et al., The Building Data Genome Project 2: Hourly energy meter data from the ASHRAE Great Energy Predictor III competition. Sci. Data. **7**, 368 (2020). https://doi.org/10.1038/s41597-020-00712-x.
9. C. Miller et al., The ASHRAE Great Energy Predictor III competition: Overview and results. Sci. Technol. Built Environ. **26,** 1427–1447 (2020). https://doi.org/10.1080/23744731.2020.1795514.
10. W. Kleiminger et al., ECO Data Set (Electricity Consumption & Occupancy), ETH Zürich Research Collection, 2016. https://www.research-collection.ethz.ch/entities/researchdata/e00ee826-4b2d-450c-8539-dc5ce14abc67.
11. ASHRAE Guideline 14-2014, Measurement of Energy, Demand, and Water Savings. American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2014.
12. A. Nouri, A. Katramiz, P. Strachan, and H. GhaffarianHoseini, Statistical methodologies for verification of building energy models: Implications for Guideline-14 metrics, in Proc. IBPSA Building Simulation 2021, 2340–2347 (2021). https://doi.org/10.26868/25222708.2021.30538.
13. S. H. Ryu and H. J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. Build. Environ. **107**, 1–9 (2016). https://doi.org/10.1016/j.buildenv.2016.06.039.
14. Y. Ding, B. Zhou, Y. Chen, and X. Ouyang, Review on occupancy detection and prediction in building. Build. Simul. **15**, 151–177 (2022). https://doi.org/10.1007/s12273-021-0813-8.
15. H. P. Das, S. Pal, and A. Bera, Machine learning for smart and energy-efficient buildings. Environ. Data Sci. **3**, e44 (2024). https://doi.org/10.1017/ eds.2024.34.