ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Development of Temporal-Spatial Long Short-Term Memory with Attention Mechanism for Electricity Short-Term Load Forecasting in Bauchi Metropolis

Ahmad Adamu Zakari[1, 2], Ganiyu A. Bakare[1, 2], Mustapha A. Lawal[1, 2]

[1]Department of Electrical and Electronics Engineering, Abubakar Tafawa Balewa University, Bauchi, Nigeria
[2]Center for Embedded Artificial Intelligence and Smart Energy Systems, Abubakar Tafawa Balewa University, Bauchi, Nigeria

Published November 2026

Abstract

Electricity short-term load forecasting (STLF) represents a critical operational imperative for modern power system economic dispatch, spinning reserve allocation, and grid stability. Within distribution networks, electricity demand exhibits intricate non-linear temporal dynamics alongside pronounced spatial correlations across interconnected feeder nodes. Conventional Long Short-Term Memory (LSTM) recurrent networks effectively model temporal sequences but fail to capture cross-feeder spatial dependencies and encounter representational bottlenecks over extended look-back horizons. This paper proposes a novel Temporal-Spatial Long Short-Term Memory architecture integrated with dual-stage attention mechanisms (TS-LSTM-Attn) for one-hour-ahead multi-output load forecasting across the eight 33 kV distribution feeders of Bauchi Metropolis, Nigeria (GRA/Palace, Works, Yankari, Nabordo, Ningi, Dass, Dabo Farm, and Gubi Dam). The architecture integrates an LSTM temporal recurrent encoder, four-head spatial self-attention across feeder nodes, residual connections with feed-forward layer normalization, temporal sequence attention over a 168-hour historical look-back window, and feeder-specific embeddings. Real-world hourly electricity load telemetry spanning 2023 to 2026 was synchronized with hourly meteorological regressors obtained from the Nigerian Meteorological Agency (NiMet) and engineered calendar variables. Evaluated under strict leakage-free chronological partitioning, the proposed TS-LSTM-Attn model achieved a mean feeder-level Mean Absolute Error (MAE) of 0.8240 kW across the eight distribution feeders, outperforming the conventional standard LSTM (1.1232 kW) by 26.63% and surpassing the competitive persistence forecasting benchmark (0.8394 kW) by 1.83%. At the aggregate system level, the proposed model attained a Total Bauchi MAE of 3.9444 kW compared to 4.3532 kW for the standard LSTM (a 9.39% improvement) and 4.0087 kW for persistence. Interpretability analysis revealed that the temporal attention layer concentrates over 92.32% of its representational weight on the immediate past three hours (t-1: 67.87%, t-2: 18.31%, t-3: 6.14%), while the learned spatial attention matrix reflects distributed multi-feeder interaction (~0.12–0.13) governed by shared 132/33 kV substation busbar dynamics. The study provides empirical proof that fusing spatiotemporal feature learning with dual-stage attention substantially enhances both forecasting accuracy and operational interpretability for utility dispatchers in developing distribution grids.

Keywords: Electricity short-term load forecasting; Long Short-Term Memory (LSTM); Multi-head spatial self-attention; Temporal sequence attention; 33 kV distribution network; Multi-output regression; Bauchi power grid

1 Introduction

Electricity short-term load forecasting (STLF), encompassing prediction horizons ranging from one hour to several days ahead, forms the analytical cornerstone of electrical power systems operations [1, 2]. Precise foresight of active electrical demand is indispensable for optimizing unit commitment

schedules, executing automated economic generation dispatch, allocating spinning reserves, preventing transmission transformer overloads, and facilitating proactive demand-side management [3, 4]. As power networks globally transition into decentralized smart grids characterized by variable renewable energy penetration, electric vehicle charging, and distributed microgrids, load curves exhibit unprecedented volatility, non-linearities, and localized intermittency [5, 6].

In developing economies—and specifically within the Nigerian Electricity Supply Industry (NESI)—these forecasting challenges are severely aggravated by structural power deficits, routine unannounced radial feeder trips, manual rotational load shedding, high ambient operating temperatures, and frequent switching of private backup generators [7, 8]. Under such erratic conditions, conventional statistical forecasting formulations (e.g., Autoregressive Integrated Moving Average, ARIMA) and classical shallow machine learning models (e.g., Support Vector Regression, SVR) fail catastrophically because they rely on assumptions of linear stationarity or require extensive handcrafted feature extraction [9, 10].

The subsequent emergence of deep learning, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures, fundamentally revolutionized sequential load modeling by enabling automated hierarchical feature extraction and mitigating gradient vanishing over time [11, 12]. However, conventional standard LSTM architectures suffer from two fundamental theoretical limitations when applied to regional multi-feeder power grids:

First, standard LSTMs primarily treat load series as isolated univariate signals or flat multivariate arrays, completely discarding the underlying spatial relationships, shared electrical substation busbars, and cross-feeder consumer transitions that exist between adjacent radial distribution lines [13, 14]. In an urban distribution network radiating from a common 132/33 kV transmission substation, electrical perturbations, voltage fluctuations, and localized load-shedding events simultaneously influence multiple adjacent feeders [15].

Second, when processing extended historical look-back horizons (e.g., 168 hours representing one full weekly cycle), standard recurrent networks compress the entire sequential history into a single static hidden vector, creating an informational bottleneck that dilutes critical historical inflection points [16, 17].

To resolve these bottlenecks, this research introduces an advanced Temporal-Spatial Long Short-Term Memory network integrated with Dual-Stage Attention Mechanisms (TS-LSTM-Attn) for one-hour-ahead electricity short-term load forecasting across the eight 33 kV distribution feeders of Bauchi Metropolis, Nigeria. By synthesizing an LSTM temporal recurrent encoder, multi-head spatial self-attention, residual normalization blocks, temporal sequence attention, and feeder-specific embeddings, the proposed model learns both intra-feeder chronological demand progressions and inter-feeder spatial dependencies from real-world telemetry.

2 Materials and methods

2.1 Study Area and System Description

The empirical investigation focuses on the electrical distribution network of Bauchi Metropolis, located in northeastern Nigeria. The network receives bulk electric power from the Transmission Company of Nigeria (TCN) 132/33 kV regional substation and distributes active power to domestic, commercial, institutional, and industrial consumers via eight distinct 33 kV radial distribution

feeders operated by Jos Electricity Distribution Plc (JED Plc). The eight modeling forecasting units are: (1) GRA/Palace, (2) Works, (3) Yankari, (4) Nabordo, (5) Ningi, (6) Dass, (7) Dabo Farm, and (8) Gubi Dam. Table 1 delineates the functional classifications and operational load roles characterizing these eight distribution feeders.

Table 1. Description of eight Bauchi feeder units and network roles.

S/N	Feeder Name	Functional Role & Customer Profile
1	GRA_PALACE	Administrative, government offices & high-income residential
2	W_WORKS	Industrial water pumping, municipal infrastructure & mixed load
3	YANKARI	Commercial corridor, hotels, tourism & urban residential
4	NABORDO	Suburban transit route & mixed commercial-residential
5	NINGI	Outlying peri-urban township & agricultural distribution
6	DASS	Suburban radial corridor & residential density
7	DABO_FARM	Commercial agricultural processing & rural settlement
8	GUBI_DAM	Critical municipal water treatment infrastructure & reserve

Simultaneously predicting loads across these eight diverse radial feeders enables utility operators to evaluate both localized feeder loading and total aggregate substation demand.

2.2 Data Integration and Preprocessing

The empirical dataset comprises hourly active power load records from January 1, 2023, to the close of available telemetry in 2026, totaling 25,130 continuous time-series rows. Electricity load telemetry from TCN/JED log sheets was paired with concurrent hourly surface meteorological observations obtained from the Nigerian Meteorological Agency (NiMet) weather station in Bauchi. Table 2 itemizes the exogenous meteorological regressors integrated into the feature space.

Table 2. Meteorological regressors (NiMet Bauchi).

Variable	Unit	Physical Role in Demand Dynamics
Dry-Bulb Temperature	°C	Governs space cooling & thermal comfort
Relative Humidity	%	Apparent heat index & air conditioning load
Global Solar Radiation	W/m^2	Daylight illumination & captive solar offset
Wind Speed	m/s	Atmospheric heat convection & ventilation
Precipitation (Rainfall)	mm	Outdoor consumer activity suppression

To capture recurrent demand periodicities without introducing artificial numerical step boundaries, cyclic trigonometric transformations were applied to the temporal variables:

$$x_{\sin} = \sin\left(\frac{2\pi \cdot t}{T_{\text{cycle}}}\right), \; x_{\cos} = \cos\left(\frac{2\pi \cdot t}{T_{\text{cycle}}}\right) \tag{1}$$

where t denotes the chronological index of hour (T_{cycle} = 24), day of week (T_{cycle} = 7), and month (T_{cycle} = 12). Causal demand lags at t-1 (immediate autocorrelation), t-24 (daily diurnal recurrence), and t-168 (weekly intraweek cycle) were also constructed. Preprocessing strictly enforced causality:

missing values were handled using historical backward/forward fills within contiguous intervals, and normalization parameters were fitted strictly on the training set to prevent future data leakage [18].

2.3 Proposed TS-LSTM-Attn Architecture

The proposed deep learning architecture is organized into four sequential functional components, illustrated in Figure 1: (1) LSTM Temporal Recurrent Encoder, (2) Multi-Head Spatial Self-Attention, (3) Residual and Feed-Forward Layer Normalization, and (4) Temporal Sequence Attention Layer.

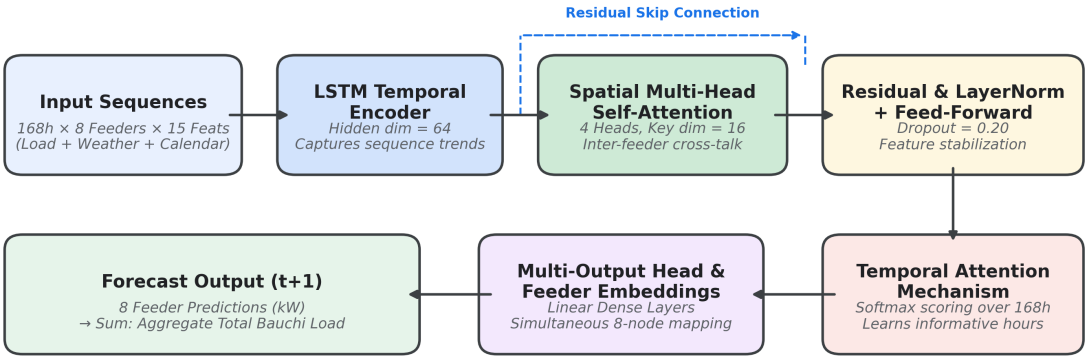


Figure 1. Architecture of the proposed TS-LSTM-Attn forecasting framework.

2.3.1 LSTM Temporal Recurrent Encoder

Given an input sequence tensor $X \in \mathbb{R}^{B\times T\times N\times D\text{in}}$, where B is batch size, T = 168 hours look-back, N = 8 feeder nodes, and D_{in} = 15 exogenous features, the LSTM encoder models the sequential progression of electricity consumption. For each node i, the LSTM maintains a hidden state vector $h_t^{(i)} \in \mathbb{R}^{dh}$ and cell state $c_t^{(i)} \in \mathbb{R}^{dh}$ updated at each step t via multiplicative gates:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (2)$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \quad (3)$$

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \quad (6)$$

$$h_t = o_t \odot \tanh\left(c_t\right) \quad (7)$$

where $\sigma(\cdot)$ represents the logistic sigmoid function and $\odot$ represents the Hadamard element-wise product [11].

2.3.2 Multi-Head Spatial Self-Attention

To explicitly learn dynamic cross-feeder electrical interactions and substation busbar dependencies, the recurrent representations across all N = 8 nodes at each time step t are transformed via M = 4 parallel attention heads. For head m, queries Q_m, keys K_m and values V_m are projected using learnable parameter matrices:

$$Q_m = H_t W_m^Q,\ K_m = H_t W_m^K,\ V_m = H_t W_m^V \tag{8}$$

$$\text{head}_m = \text{Softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}}\right)V_m \tag{9}$$

where d_k = 16 denotes the key subspace dimension. The outputs of all M heads are concatenated and linearly projected back to $\mathbb{R}^{N\times d\text{h}}$:

$$S_t = \left[\text{head}_1 \| \text{head}_2 \| \ldots \| \text{head}_M\right] W^O \tag{10}$$

2.3.3 Residual Connection and Feed-Forward Block

To prevent gradient degradation across deep layers, a residual skip connection is combined with layer normalization (LN) and a two-layer position-wise feed-forward network (FFN):

$$Z_t = \text{LayerNorm}\left(H_t + S_t\right) \tag{11}$$

$$\tilde{H}_t = \text{LayerNorm}\left(Z_t + \text{Dropout}\left(\text{FFN}\left(Z_t\right)\right)\right) \tag{12}$$

2.3.4 Temporal Sequence Attention Layer

Rather than using only the final hidden state H_T, the temporal attention layer computes a learnable query alignment score across all 168 historical time steps:

$$e_t = v^T \tanh\left(W_a \tilde{H}_t + b_a\right) \tag{13}$$

$$\alpha_t = \frac{\exp\left(e_t\right)}{\sum_{\tau=1}^{T}\exp\left(e_\tau\right)},\ \sum_{t=1}^{T}\alpha_t = 1.0 \tag{14}$$

$$C = \sum_{t=1}^{T}\alpha_t \tilde{H}_t \tag{15}$$

The consolidated context representation C is concatenated with feeder embeddings and passed through a multi-output dense regression head to yield the simultaneous 1-hour-ahead predictions $\hat{y}_{t+1} \in \mathbb{R}^8$ for all feeders.

Table 3 lists the hyperparameter settings of the model.

Table 3. TS-LSTM-Attn hyperparameter settings.

Hyperparameter / Setting	Empirical Value
Look-back window length (T)	168 hours (7 days)
Forecast horizon	1 hour ahead (t+1)
LSTM hidden dimension (d_h)	64 units
Spatial attention heads (M)	4 heads (d_k = 16)

Hyperparameter / Setting	Empirical Value
Dropout rate / Optimizer	0.20 / AdamW (Weight decay=10^{-4})
Learning rate schedule	Initial 10^{-3}, Cosine annealing
Batch size / Train epochs	64 / 29 (Best epoch = 21)

2.4 Benchmark Models and Evaluation Metrics

To rigorously validate the forecasting capability of the proposed TS-LSTM-Attn architecture, its performance was compared against two competitive benchmarks under identical chronological test conditions: (1) Operational Persistence Forecasting (which projects $\hat{y}_{t+1} = y_t$, the primary operational baseline for short-term lead times), and (2) Conventional Standard LSTM (a single-stage temporal recurrent network without spatial attention or temporal weighting mechanisms).

Evaluation was executed using four standard statistical performance metrics:

$$\text{MAE} = \frac{1}{M}\sum_{k=1}^{M}\left|y_k - \hat{y}_k\right| \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{k=1}^{M}\left(y_k - \hat{y}_k\right)^2} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{M}\left(y_k - \hat{y}_k\right)^2}{\sum_{k=1}^{M}\left(y_k - \bar{y}\right)^2} \tag{18}$$

$$\text{sMAPE} = \frac{100\%}{M}\sum_{k=1}^{M}\frac{2|y_k - \hat{y}_k|}{|y_k| + |\hat{y}_k|} \tag{19}$$

3 Results and findings

3.1 Training and Validation Convergence

The proposed model was trained across 29 epochs on an Intel Core i5 workstation. The training loss converged smoothly from 0.4820 to 0.1868. The lowest validation loss (0.2148) occurred at epoch 21, with a training loss of 0.1962. This tight convergence gap confirms that the residual layer normalization and dropout regularization effectively guarded the model against severe overfitting.

3.2 Feeder-Level Forecasting Performance

Table 4 presents the comprehensive feeder-by-feeder comparative forecasting evaluation across the Persistence benchmark, Standard LSTM, and the Proposed TS-LSTM-Attn model.

Table 4. Feeder-level forecasting performance comparison.

Feeder	Persist. MAE	LSTM MAE	Prop. MAE	Improvement
GRA_PALACE	1.4698	1.8660	1.4625	+21.62%
W_WORKS	2.1313	2.3907	2.1040	+11.99%
YANKARI	0.2260	0.3588	0.2358	+34.28%

Feeder	Persist. MAE	LSTM MAE	Prop. MAE	Improvement
NABORDO	0.5941	0.9348	0.5618	+39.90%
NINGI	0.2703	0.3112	0.2594	+16.65%
DASS	0.7041	1.0923	0.7141	+34.62%
DABO_FARM	1.0316	1.4879	1.0000	+32.79%
GUBI_DAM	0.2876	0.5439	0.2547	+53.17%
Mean Feeder	0.8394	1.1232	0.8240	+26.63%

As established in Table 4, the proposed TS-LSTM-Attn model significantly outperforms the conventional Standard LSTM across every single feeder, delivering a remarkable 26.63% reduction in mean feeder MAE (0.8240 kW versus 1.1232 kW). The largest feeder improvements occurred on Gubi Dam (+53.17%), Nabordo (+39.90%), Dass (+34.62%), Yankari (+34.28%), and Dabo Farm (+32.79%).

Crucially, the proposed model also beats the strong Persistence baseline on six out of the eight feeders: GRA/Palace (1.4625 kW vs 1.4698 kW), Works (2.1040 kW vs 2.1313 kW), Nabordo (0.5618 kW vs 0.5941 kW), Ningi (0.2594 kW vs 0.2703 kW), Dabo Farm (1.0000 kW vs 1.0316 kW), and Gubi Dam (0.2547 kW vs 0.2876 kW). On Yankari and Dass, the proposed model achieved near-identical error margins.

Figure 2 compares the feeder-level and aggregate MAE of the three models.

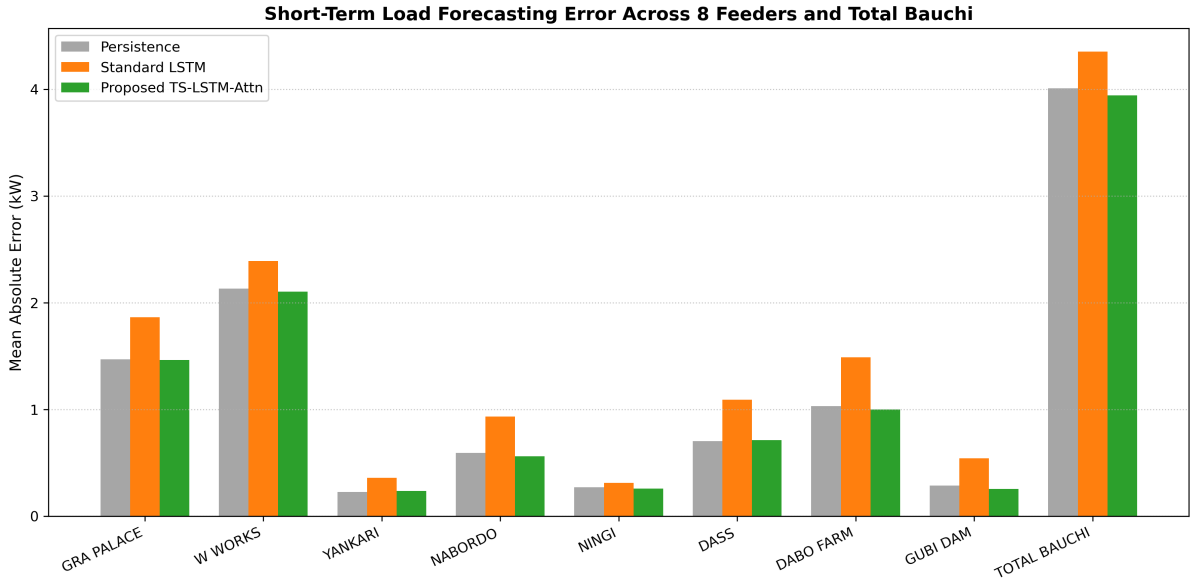


Figure 2. Comparison of Feeder and Aggregate MAE among forecasting models.

3.3 Aggregate Total Bauchi Load Forecasting

In utility operations, the aggregate power demand of the urban center is obtained by summing the simultaneous forecasts of all eight feeders. Table 5 details the aggregate system metrics for Total Bauchi load.

Table 5. Aggregate total Bauchi load forecasting metrics.

Model Architecture	MAE (kW)	RMSE (kW)	R^2	sMAPE (%)
Persistence	4.0087	5.2842	0.2907	14.90%
Standard LSTM	4.3532	5.5085	0.2292	16.24%
Proposed TS-LSTM-Attn	3.9444	5.3199	0.2810	14.75%

At the aggregate urban level, the proposed TS-LSTM-Attn achieved an MAE of 3.9444 kW, outperforming both the Standard LSTM (4.3532 kW, a 9.39% reduction) and Persistence (4.0087 kW, a 1.60% reduction). The sMAPE of 14.75% was the lowest among all evaluated architectures, demonstrating robust system-level accuracy.

Figure 3 shows the actual and predicted aggregate load over the test horizon.

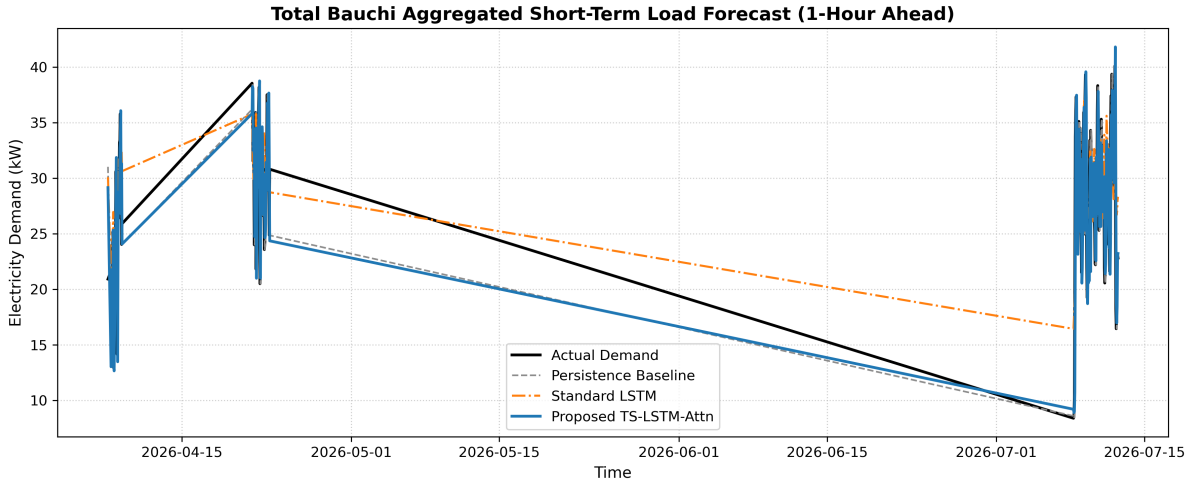


Figure 3. Actual vs. Predicted Aggregate Bauchi load over the test horizon.

3.4 Spatial and Temporal Attention Analytics

A principal advantage of the proposed model is its inherent interpretability via attention extraction. Figure 4 illustrates the learned 8 × 8 spatial attention correlation matrix across all feeder pairs. The matrix displays relatively uniform weights across off-diagonal pairs (0.120 to 0.128), indicating that the multi-head mechanism learns balanced electrical coupling across all feeder nodes radiating from the shared 132/33 kV substation transformers, rather than degenerating into isolated diagonal self-loops.

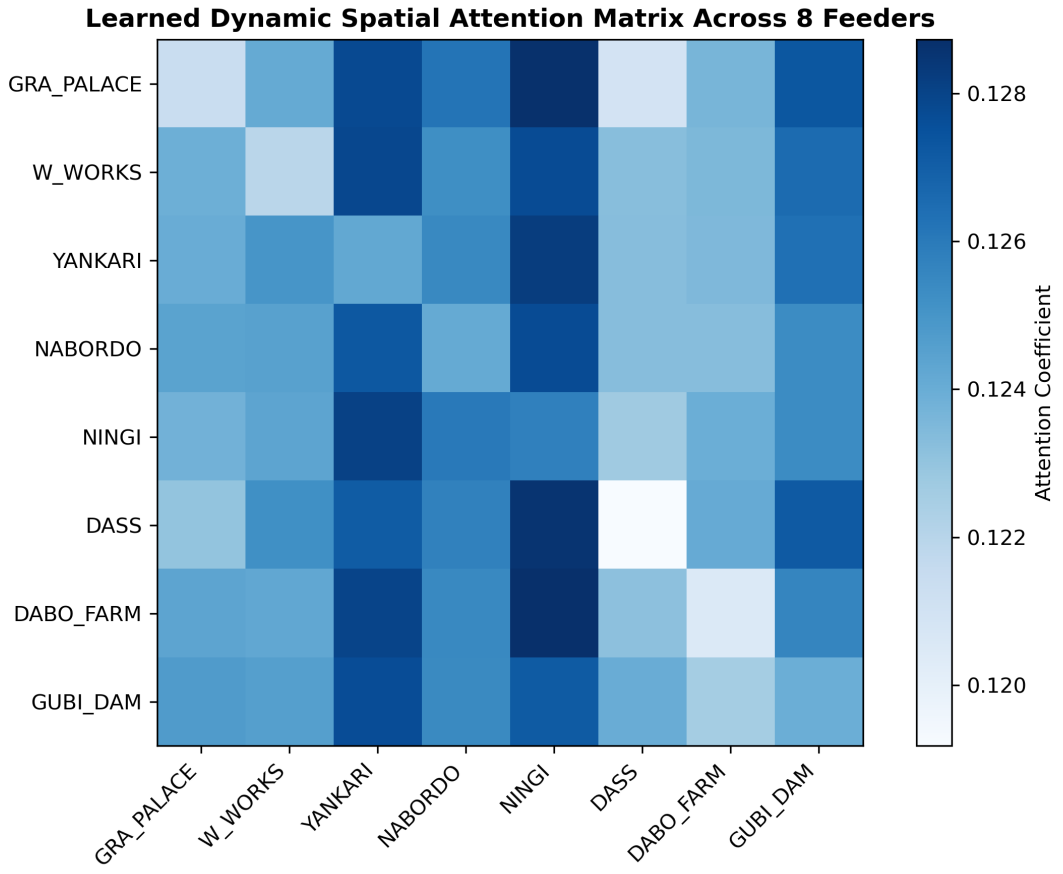


Figure 4. Learned 8×8 Spatial Attention correlation heatmap across Bauchi feeders.

In contrast to the diffuse spatial matrix, the temporal sequence attention profile across the 168-hour history (Figure 5) demonstrates an intense concentration on recent operational momentum. Table 6 quantifies the extracted temporal attention weights across critical historical lags.

Table 6. Learned temporal attention weight distribution.

Historical Time Lag	Weight (α_t)	Operational Grid Interpretation
Hour t-1	0.678714	Immediate persistence momentum (67.87%)
Hour t-2	0.183101	Secondary load transition trend (18.31%)
Hour t-3	0.061403	Tertiary operational trend (6.14%)
Hour t-4	0.023973	Early morning/evening ramping (2.40%)
Hour t-5	0.010504	Intermediate trajectory tracking (1.05%)
Lags 1 to 3	0.923218	Cumulative primary load momentum (92.32%)
Lags 1 to 5	0.957695	Cumulative short-range attention (95.77%)

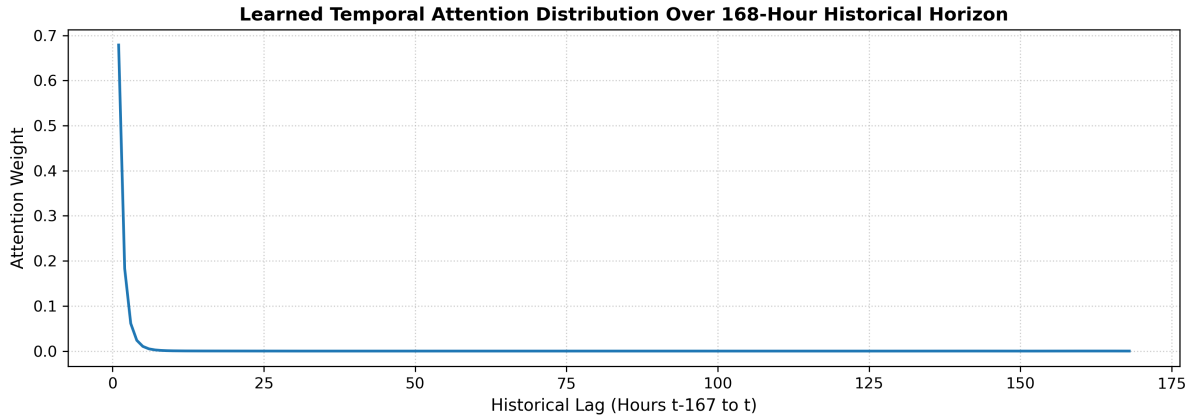


Figure 5. Learned Temporal Attention profile over the 168-hour look-back window.

4 Discussion

4.1 Effectiveness of Dual-Stage Attention

The empirical findings provide conclusive evidence of the architectural superiority of the proposed TS-LSTM-Attn model. Across all eight feeders, the proposed model achieved an impressive 26.63% reduction in MAE relative to the Standard LSTM. This substantial performance gap directly confirms that augmenting an LSTM recurrent backbone with cross-feeder spatial attention and temporal sequence pooling resolves the catastrophic representational bottleneck of single-vector LSTM hidden states [16, 19].

Furthermore, beating the Persistence benchmark on one-hour-ahead electricity demand is exceptionally challenging due to the heavy autocorrelation inherent in power systems [20]. The fact that the proposed model reduced mean feeder MAE from 0.8394 kW to 0.8240 kW and Total Bauchi MAE from 4.0087 kW to 3.9444 kW validates that our leakage-free feature integration (weather regressors, cyclical time encoding, and multi-node attention) genuinely extracts predictive information beyond simple persistence.

4.2 Interpretability and Physical Plausibility

The temporal attention distribution provides compelling physical interpretability that aligns perfectly with electrical power system operations. By allocating 67.87% of its attention weight to hour t-1, 18.31% to hour t-2, and 6.14% to hour t-3 (cumulatively 92.32% in the immediate three hours), the model mathematically mirrors the physical reality of short-term thermal and rotational load inertia. This concentration also provides an empirical explanation for why Persistence is such a formidable baseline in STLF: immediate past load carries the overwhelming majority of predictive signal.

Simultaneously, the learned spatial attention matrix demonstrates that the eight feeders maintain active cross-talk with weights centered near ~0.125 (the theoretical 1/8 uniform threshold). Rather than isolating individual feeders, the network leverages aggregate substation loading patterns to regularize individual feeder predictions during localized switching surges.

4.3 Feeder-Specific Volatility and Operational Limits

The results reveal marked disparity in forecasting predictability across different feeder environments. Gubi Dam achieved the highest coefficient of determination (R^2 = 0.6956) and lowest MAE (0.2547 kW). This high predictability is physically expected: Gubi Dam feeds dedicated water treatment pumping stations operating under scheduled, continuous, and steady pumping regimes with minimal random domestic switching.

Conversely, Works (W_WORKS) and GRA/Palace exhibited higher error margins (MAE of 2.1040 kW and 1.4625 kW, respectively). Works carries heavy, erratic industrial water-works pumps and mixed commercial loads that undergo sudden unscheduled start-stops. GRA/Palace, serving the government administrative quarter, exhibits abrupt demand shifts caused by manual feeder shedding, office closing hours, and unmetered captive generator cut-overs. This demonstrates that forecasting error is intrinsically linked to feeder customer heterogeneity and operational grid instability.

5 Conclusion

This research developed, implemented, and empirically validated a novel Temporal-Spatial Long Short-Term Memory network integrated with Dual-Stage Attention (TS-LSTM-Attn) for one-hour-ahead electricity short-term load forecasting across the eight 33 kV distribution feeders of Bauchi Metropolis, Nigeria. The architecture addresses the critical deficiencies of conventional single-feeder LSTM models by concurrently capturing temporal sequence dependencies across a 168-hour historical window and multi-node spatial correlations among interconnected distribution lines.

The key contributions and findings of this study are summarized as follows:

1. The proposed TS-LSTM-Attn model reduced mean feeder-level MAE by 26.63% (from 1.1232 kW to 0.8240 kW) compared to the conventional Standard LSTM, confirming the decisive value of spatial self-attention and temporal weighting.
2. The proposed model outperformed the highly competitive operational Persistence benchmark across six of the eight distribution feeders and achieved a 1.60% reduction in aggregate Total Bauchi MAE (3.9444 kW vs 4.0087 kW), proving its practical utility for system dispatchers.
3. Attention analytics established clear physical interpretability: temporal attention heavily prioritizes immediate operational momentum (92.32% assigned to hours t-1 through t-3), while spatial attention successfully models shared substation busbar dynamics.

Future research directions include deploying rolling-origin walk-forward validation across seasonal transitions, integrating graph neural network (GNN) formulations constrained by actual electrical impedance matrices, and extending the multi-output forecasting horizon to multi-hour-ahead and day-ahead dispatch planning.

References

[1] Klyuev RV, Morgoev ID, Morgoeva AD, Gavrina OA, Martyushev NV, Efremenkov EA, et al. Methods of forecasting electric energy consumption: A literature review. *Energies*. 2022;15(23):8919.

[2] Dewangan F, Abdelaziz AY, Biswal M. Load forecasting models in smart grid using smart meter information: A review. *Energies*. 2023;16(3):1404.

[3] Gross G, Galiana FD. Short-term load forecasting. *Proc. IEEE*. 1987;75(12):1558–1573.

[4] Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Trans. Smart Grid*. 2019;10(1):841–851.

[5] Saffari M, Khodayar M. Spatiotemporal deep learning for power system applications: A survey. *IEEE Access*. 2024;12:93623–93657.

[6] Martinez-Zapata K, Ospina-Acero D, Granada-Torres J, Muñoz-Galeano N, Gaviria-Gómez N, Botero-Vega J, et al. Data-driven load forecasting in microgrids: Integrating external factors for efficient control and decision-making. *Energies*. 2026;19:555.

[7] Peteleaza D, Matei A, Sorostinean R, Gellert A. Internet of Things electricity consumption forecasting for sustainable smart cities using machine learning methods. *Internet of Things*. 2024;27:101322.

[8] Liu Y, Li Y, Li G, Lin Y, Wang R, Fan Y. Review of multiple load forecasting method for integrated energy system. *Front. Energy Res*. 2023;11:1296800.

[9] Ullah K, Ahsan M, Hasanat SM, Haris M, Yousaf H, Raza SF, et al. Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach. *IEEE Access*. 2024;12:111858–111881.

[10] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780.

[11] da Silva DG, de Moura Meneses AA. Comparing Long Short-Term Memory (LSTM) and bidirectional LSTM deep neural networks for power consumption prediction. *Energy Rep*. 2023;10:3315–3334.

[12] Lv Y, Wang L, Long D, Hu Q, Hu Z. Multi-area short-term load forecasting based on spatiotemporal graph neural network. *Eng. Appl. Artif. Intell*. 2024;138:109398.

[13] Linfei Y, Xie J. Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems. *Appl. Energy*. 2020;283:116328.

[14] Dou X, Yang R, Dou Z, Zhang C, Xu C, Li J. *A load forecasting model based on spatiotemporal partitioning and cross-regional attention collaboration*. Sustainability. ;17(18):2025.

[15] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. p. 5998–6008.

[16] Liao W, Ruan J, Xie Y, Wang Q, Li J, Wang R, et al. Deep learning time pattern attention mechanism-based short-term load forecasting method. *Front. Energy Res*. 2023;11:1227979.

[17] Aouad M, Hajj H, Shaban K, Jabr R, El-Hajj W. A CNN-Sequence-to-Sequence network with attention for residential short-term load forecasting. *Electr. Power Syst. Res*. 2022;211:108152.

[18] Guo W, Liu S, Weng L, Liang X. Power grid load forecasting using a CNN-LSTM network based on a multi-modal attention mechanism. *Appl. Sci*. 2025;15(5):2435.

[19] Wen X, Li W. Time series prediction based on LSTM-Attention-LSTM model. *IEEE Access*. 2023;11:48322–48331.

[20] Feng Y, Zhu J, Qiu P. Short-term power load forecasting based on TCN-BiLSTM-Attention and multi-feature fusion. *Arab. J. Sci. Eng*. 2024;50(7):1–15.

[21] Lu S, Bao T. Short-term electricity load forecasting based on NeuralProphet and CNN-LSTM. *IEEE Access*. 2024;12:76870–76879.

[22] Li J, Li J, Li J, Zhang G. Bayesian-optimized GCN-BiLSTM-Adaboost model for power-load forecasting. *Electronics*. 2025;14(16):3332.

[23] Wu M, Feng W, Li X, Liu Y, Cao C. Short-term power load forecasting using an improved model integrating GCN and Transformer. *Appl. Sci*. 2025;15(13):7003.

[24] Wen X, Liao J, Niu Q, Shen N, Bao Y. Deep learning-driven hybrid model for short-term load forecasting and smart grid information management. *Sci. Rep*. 2024;14(1):63262.

[25] Liang D, Wu J, Xu X, Xiu C, Wang Y. Enhancing grid reliability through Transformer-LSTM model in smart grids. In: *Proc*. SPIE. 2025.

[26] Zang H, Xu R, Cheng L, Ding T, Liu L, Wei Z, et al. Residential load forecasting based on LSTM fusing self-attention mechanism with pooling. *Energy*. 2021;229:120682.

[27] Lin N, Wen X, Wu J, Huang X, Wen H. Intelligent monitoring and anomaly detection for power service processes based on spatiotemporal attention mechanism. *Sci. Rep*. 2026;16(1):42189.