Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Comparative Evaluation of LSTM, BiLSTM, and a Hybrid CNN-BiLSTM-Attention Model for Short-Term and Day-Ahead Electricity Load Forecasting

Sadeeq Ullah*
Department of Computer Science, Bacha Khan University Charsadda
Email: sadeequllahsadeequllah47@gmail.com

Asad Ullah*
Department of Computer Science, Bacha Khan University Charsadda
Email: asadullahofficial95@gmail.com

Atta Ur Rahman*
Lecturer, Department of Computer Science, Bacha Khan University Charsadda
Email: atta@bkuc.edu.pk

Naeem Jan
Lecturer, Department of Computer Science, Bacha Khan University Charsadda
Email: naeemjan@bkuc.edu.pk

Muhammad Shahid
Lecturer, ICS/IT, The University of Agriculture, Peshawar
Email: mshahidkhan32@gmail.com

ABSTRACT

Reliable electricity load forecasting is essential for grid stability, generation scheduling, and demand-supply balancing, particularly in regions such as Pakistan that face chronic supply shortfalls. Building on a base study that compared Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) models against ARIMA and SARIMA on a univariate prosumer dataset, this paper extends the comparison in three directions: it evaluates a proposed Hybrid CNN-BiLSTM-Attention architecture alongside the LSTM and Bi-LSTM baselines, it replaces the single date time feature of the base study with a fifteen-column multivariate feature set (cyclical calendar encodings, lag features at 1, 24, and 168 hours, and 24-hour rolling statistics), and it evaluates every architecture at both a 1-hour-ahead and a 24-hour-ahead forecasting horizon rather than one-step-ahead alone. All three models were trained and tested under identical conditions on the public PJM East (PJME_MW) hourly electricity load series (145,224 usable hourly records, January 2002 to August 2018) using a 168-hour lookback window, an 80/10/10 chronological train/validation/test split, the Adam optimizer, and early stopping. On the held-out test set, Bi-LSTM produced the lowest error for 1-hour-ahead forecasting (RMSE = 528.17 MW, MAE = 402.62 MW), while LSTM was marginally the most accurate model for 24-hour-ahead forecasting (RMSE = 2014.78 MW, MAE = 1475.48 MW),

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

narrowly ahead of Bi-LSTM (RMSE = 2033.93 MW). The Hybrid CNN-BiLSTM-Attention model recorded the highest absolute error at both horizons (RMSE = 2658.79 MW at h = 1; RMSE = 2467.19 MW at h = 24), but its error changed by only -7.2% between the two horizons, compared with +103.4% for LSTM and +285.1% for Bi-LSTM, indicating markedly better cross-horizon stability. Attention-weight analysis showed the hybrid model consistently assigning the largest weight to the most recent hour and a secondary peak at the 168-hour (one-week) lag, giving a transparent, interpretable account of which historical hours' drive each forecast. The results suggest that, under the tested configuration, richly engineered multivariate features contribute more to forecasting accuracy than additional architectural complexity, while attention-based hybrid models trade some absolute accuracy for stability and interpretability across forecast horizons.

Keywords: Electricity Load Forecasting, Long Short-Term Memory, Bidirectional Lstm, Convolutional Neural Network, Attention Mechanism, Time Series Forecasting, Deep Learning

Introduction

A reliable and stable electrical grid is critical to the economy, local industry and convenience in everyday life. Climbing electricity demand is happening around the world because information technology is consuming more energy, cities are growing, and the population is increasing. This places a significant stress on power generation and power transmission lines. Supply-demand imbalance is an ever-present problem in Pakistan, a developing country. This results in regular power cuts and unstable grid, impacting businesses and communities [1]. There are few practical tools to balance supply and eliminate blackouts that can be used immediately, and accurate demand forecasting is one of these.Building new power plants takes years, and accurate demand forecasting is one of the few tools right now available to balance supply and prevent blackouts.

Electricity usage is a highly non-linear time series. It varies according to weather, calendar (hour of day, day of the week, holiday etc.), seasonal changes, and user behavior [2]. Commonly used statistical models such as ARIMA and SARIMA are suitable for linear and stationary data. But they have difficulty modelling the non-linear patterns that are characteristic of real-world electricity loads [3]. This mismatch frequently leads to sub-optimal operational dispatch decisions, and to unnecessary high reserve margins.

The approach of deep learning is a new way to tackle load forecasting. Recurrent Neural Networks (RNNs), especially the Long Short-Term Memory network and its bidirectional variant Bi-LSTM can be designed to learn long-range time dependencies [4]. They consistently beat classical statistical models on time series tasks. For instance, Alizadegan and colleagues compared LSTM, Bi-LSTM, ARIMA, and SARIMA using the Kaggle Enefit prosumer dataset. Their results showed that Bi-LSTM had the lowest error, cutting RMSE by 5.35 percent compared to LSTM, 46.08 percent compared to ARIMA, and 50.6 percent compared to SARIMA. Their MAE drops followed a similar pattern [5]. Even so, that study relied on a single date-time feature, looked only at one-step-ahead forecasts, and recommended testing hybrid

Online ISSN 3007-3197
Print ISSN 3007-3189

LSTM architectures in the future.

This paper takes up that exact recommendation. Three specific gaps in the earlier study motivate this work. First, a standard LSTM or Bi-LSTM does not point out which past time steps actually drive the forecast. Pairing a convolutional feature extractor with an attention mechanism can improve representational capacity while making model decisions easier to interpret. Second, feeding a model only a single timestamp ignores the predictive value hidden in past load values and rolling statistics. Third, testing models only on a one-step-ahead horizon misses how they perform over the longer 24-hour day-ahead horizon needed for actual grid scheduling.

To address these gaps, this study introduces four changes. It implements a Hybrid CNN-BiLSTM-Attention model that builds on the baseline architectures. It builds a fifteen-column multivariate feature set using cyclical time encodings, lag features, and rolling statistics instead of a single date-time column. It evaluates every model at both one-hour and 24-hour horizons under identical training conditions. Finally, it measures the cross-horizon stability of each architecture by tracking how much the error grows between the one-hour and 24-hour marks.

Because multi-step forecasting and rich features require a denser dataset than the roughly 15,000 observations in the Enefit data, this study uses the publicly available PJM East hourly load series, which spans more than 16 years. Because of this, the absolute error numbers cannot be compared directly to the base paper. Instead, this work serves as an independent test of these architectures using a hybrid attention model, multivariate features, and a multi-step horizon on an industrial-scale dataset.

The rest of the paper covers related work in Section 2, the dataset and model training details in Section 3, experimental results in Section 4, and conclusions with future directions in Section 5.

Literature Review

LSTMs, in particular, have been shown to be useful in many forecasting applications. They have been used in combination with various optimization algorithms for electricity price prediction, in stacked architectures for outsmarting common benchmarks for corporate demand-planning, and even for the demand of tourism and financial markets [6]. If one focuses on the short time horizon of electricity consumption, then it is shown that the bidirectional approach is superior to the standard LSTM approach in every study conducted so far. Other studies on solar energy forecasting indicated that decomposing the solar energy data combined with the Bi-LSTMs further reduced prediction errors than the basic baseline models [4].

The research work presented here is based on the work performed by Alizadegan et al., which they applied different statistical models such as ARIMA and SARIMA as well as machine learning and deep learning techniques to the Enefit prosumer dataset. They noticed that the accuracy of Bi-LSTM was higher and the root mean square errors were much lower than the classical statistical models. Table 1 breaks down this study alongside seven other recent papers that highlight the exact gaps this research aims to fill.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 1. Summary of related work on energy consumption forecasting.

Study	Dataset	Method	Key finding	Identified gap
Alizadegan, Rashidi Malki [5]	Enefit prosumer dataset (Kaggle)	LSTM, Bi-LSTM vs. ARIMA, SARIMA	Bi-LSTM best; RMSE 46.08%/50.6% lower than ARIMA/SARIMA	Single date-time feature; one-step-ahead only; no hybrid model
Shapi, Ramli [7]	Smart building energy data (Malaysia)	Multiple ML regressors	Random Forest / SVM performed best among tested ML models	Building-scale only; no deep sequence models
Khalil, McGough [8]	Systematic review, multiple studies	Review of ML, DL, statistical methods	Hybrid models and non-linear handling identified as key gaps	No experimental implementation
Momani, Abwaini [9]	US annual grid data; IDECO daily 11 kV feeder	ARIMA vs. SARIMA comparison	ARIMA better for long-term; SARIMA better for seasonal short-term	No deep learning; limited to statistical models
Alzoubi [10]	Smart-home appliance and household dataset	Data fusion-based ML	~92% prediction accuracy	Limited to smart-home scale; no recurrent models
Wu and Wu [11]	Residential electricity consumption	CNN-BiLSTM-SA (self-attention)	Outperformed CNN-BiLSTM and BiLSTM baselines	Residential scale; no explicit multi-horizon evaluation
Bilgili and Pinar [12]	Türkiye gross electricity consumption	LSTM vs. SARIMA	LSTM MAPE 2.42%; SARIMA substantially worse	Single country; no hybrid architecture evaluated
Ravivarma, Lokeshgupta [13]	Smart micro-grid load and renewable generation	Hybrid BiLSTM-GRU	Outperforms single-architecture ML/DL models	High computational complexity

Three consistent gaps emerge from this body of work. First, few studies benchmark a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

CNN-BiLSTM-attention hybrid against standalone LSTM and Bi-LSTM baselines under identical experimental conditions. Second, most studies employ limited or univariate features as inputs instead of a systematically engineered multivariate feature set incorporating the lag values, rolling statistics and calendar encodings of the cyclical variables. Third, the evaluation is normally limited to just one forecasting horizon, while the comparative performance of architectures at short and day ahead time scales is largely ignored. The current study directly resolved these three gaps.

Methodology

Research Design

The study is conducted in a controlled experiment to get a stringent and unbiased assessment. Three different architectures are trained and tested under the same conditions: 1) Baseline Long Short-Term Memory network (LSTM), 2) Bi-directional Long Short-Term Memory (BiLSTM), and 3) the proposed hybrid CNN-BiLSTM-Attention model. All models are based on the same preprocessed data, feature set, chronologically split data, and optimization parameters. Evaluation of performance is done on two time scales, one hour ahead (h = 1) and twenty-four hours ahead (h = 24). Since all external factors are kept constant, any difference in predictive accuracy can be directly attributed to differences in the architectures. In addition, each model is given the same 168-hour "look back window," which means it looks at one full week of historical data before making a prediction. This is all the work flow starting from data ingestion till the final evaluation as shown in figure-1.

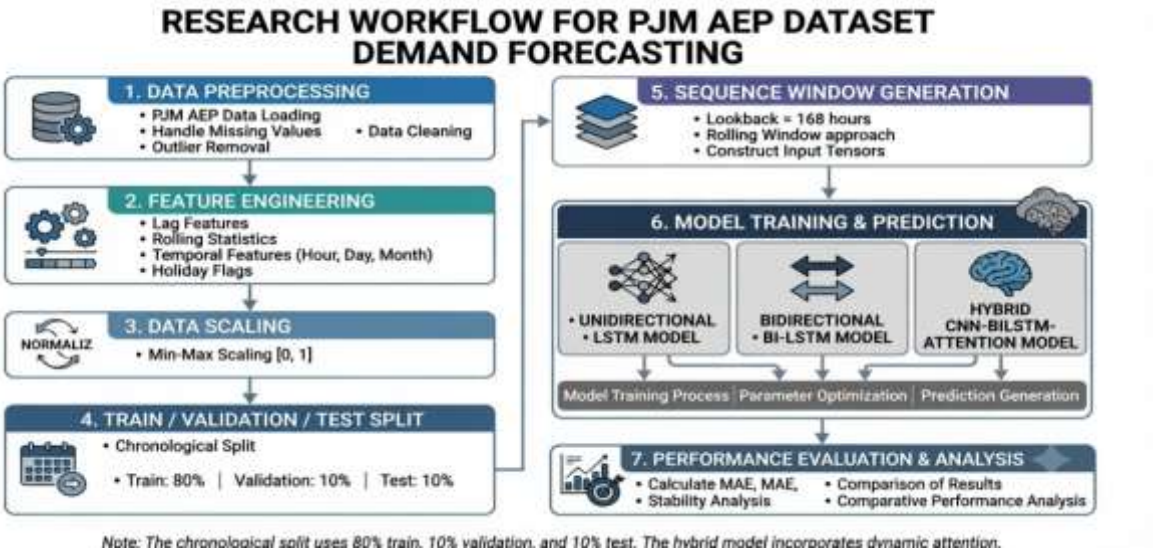


Figure 1. Research workflow for the PJM East load-forecasting comparison.

Dataset

This study replaces the Enefit prosumer dataset which was used in the base paper with the PJM East (PJME_MW) hourly electricity load series, as the PJM East series provides a much longer and purer hourly record, which is more appropriate for multi-step forecasting and feature engineering. The raw data consists of 145,366 date-time and load (MW) measurements per hour. The cleaned series now has 145,392 hourly

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

records from 1 January 2002 to 3 August 2018, with 30 short gaps (less than 6 hours) filled in using time-weighted linear interpolation. The descriptive statistics of the cleaned load series are reported in Table 2, and Figure 2 illustrates the overall trend of the load series with the monthly and hourly load distributions.

Table 2. Descriptive statistics of the cleaned hourly load series (PJM East, 2002–2018).

Statistic	Value (MW)
Count	145,392
Mean	32,078.93
Standard deviation	6,464.28
Minimum	14,544.00
25th percentile	27,571.00
Median	31,420.00
75th percentile	35,647.00
Maximum	62,009.00

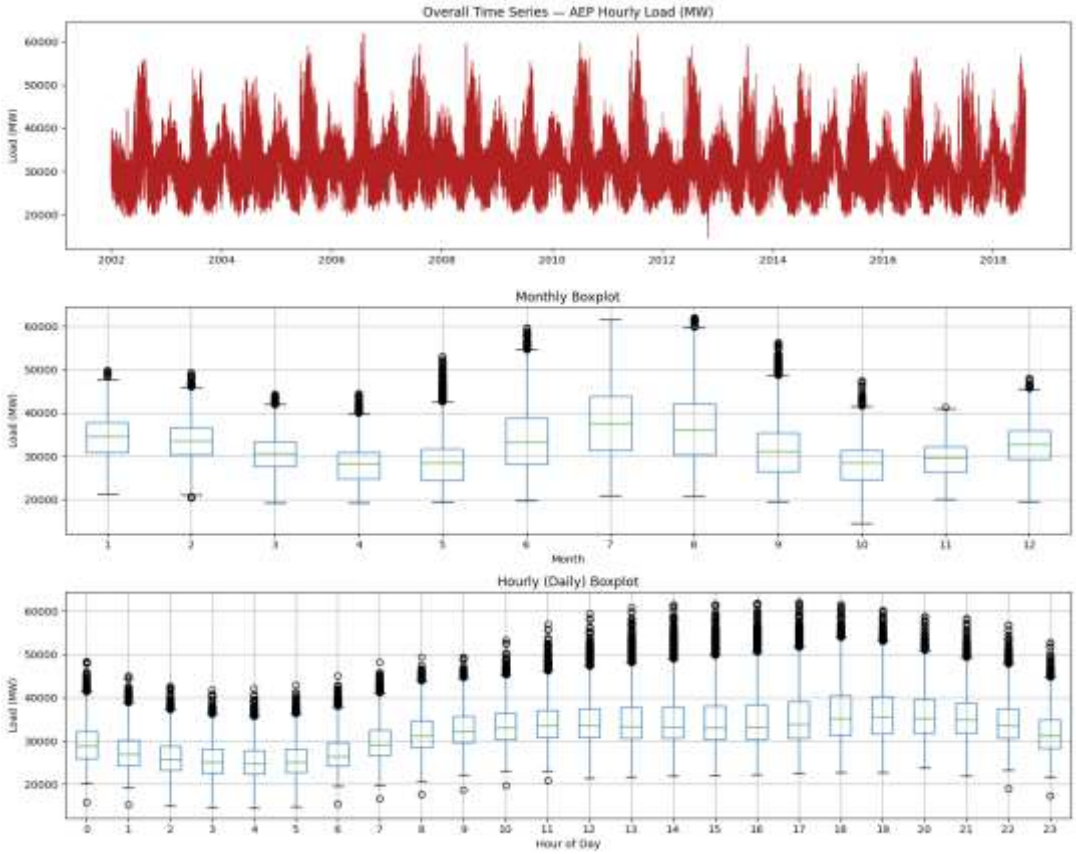


Figure 2. Overall PJM East hourly load series (top), monthly load distribution (middle), and hourly load distribution (bottom).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The PJM East series shows clear seasonal regularities as seen in Figure 2. During the summer months (June-July-August), intense cooling loads cause load peaks, and during the winter months (December-January), heating demands cause load peaks, with tapering off during the more moderate spring and autumn seasons. In addition, the intraday distribution shows a distinct morning load ramp and a flat load distribution during the afternoon and evening hours, which is characteristic of the load distribution of commercial and residential activities.

Feature Engineering

The original authors mentioned that this was a limitation of the foundational study, where only one value of date-time was used for input. This research is based on a 15-column feature set that is completely comprised of the timestamp and the past values of the load. External data sources, like weather observations, are not included. The hour of the day, day of the week, month of the year and day of the year are all circular, and these are encoded through sine and cosine transformations. This change puts hour 23 and hour 0 close together in the list, instead of far apart.

Setting the lag to 1, 24 and 168 hours allows the network to have direct access to electricity demand data from one hour, one day and one week ago. This design is designed to encode daily and weekly cycles into the inputs, rather than having to be learned by the recurrent layers, which is what most current designs do. A 24 hour rolling mean and rolling standard deviation reflects the direction and volatility of a short-term trend. The complete feature set is listed in Table 3. These lag and rolling metrics generate missing values in the first few rows that must be removed to give a final training set with 145,224 rows.

Table 3. Summary of the fifteen-column engineered feature set.

Category	Features
Target / raw	Load
Cyclical time-of-day	hour_sin, hour_cos
Cyclical day-of-week	dow_sin, dow_cos
Cyclical month-of-year	month_sin, month_cos
Cyclical day-of-year	doy_sin, doy_cos
Calendar flag	is_weekend
Lag features	lag_1h, lag_24h, lag_168h
Rolling statistics	roll_mean_24h, roll_std_24h

Data Splitting, Scaling, and Sequence Windowing

To avoid data leakage—an important prerequisite when forecasting times series—the dataset is partitioned without shuffling by time. Based on the split ratio agreed upon in foundational work, the data is split into 80% training, 10% validation, and 10% test sets. This structure results in 116,179 training rows spanning from January 8, 2002, to

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

April 10, 2015; 14,522 validation rows from April 10, 2015, to December 5, 2016; and 14,523 test rows from December 5, 2016, to August 3, 2018.

All features and target values are normalized on the training partition, without any leakage from the validation or test partitions. For the feature matrix and target variable, there are separate scaler instances, which enables the model predictions to be easily inverse transformed back to megwatts, for accurate evaluation. Lastly, sequences are shaped by inputting into a window that is 168 hours in length and 24 hours in advance. This configuration generates 116,011 training sequences, 14,354 validation sequences and 14,355 test sequences for 1 hour ahead horizon and 115,988 training sequences, 14,331 validation sequences and 14,332 test sequences for 24 hour ahead horizon.

Model Architectures

Baseline LSTM

The basic LSTM is a structure similar to the one used in the initial research. It is composed of two LSTM layers with respectively 128 and 64 units. After every recurrent layer a dropout layer with dropout rate 0.2 is introduced to prevent the overfitting that could occur and this is then passed on to a dense layer with 64 ReLU neurons, and a linear output layer with 24 units. Note that the model automatically provides a full 24-step forecast trajectory; if the evaluation is performed at the 1-hour-ahead time horizon (h = 1), then only the first element of the output vector is evaluated. This configuration consists of 128,856 trainable parameters, which use up around 503.34 KB of memory, as shown in Table 4.

Table 4. Baseline LSTM model summary (128,856 trainable parameters).

Layer (type)	Output shape	Parameters
Input	(None, 168, 15)	0
LSTM	(None, 168, 128)	73,728
Dropout	(None, 168, 128)	0
LSTM_1	(None, 64)	49,408
Dropout_1	(None, 64)	0
Dense	(None, 64)	4,160
Dense_1 (output)	(None, 24)	1,560

Baseline Bidirectional LSTM (Bi-LSTM)

The bidirectional LSTM baseline is the same as the standard LSTM baseline in terms of topology. But in the place of each unidirectional LSTM layer, a bidirectional LSTM layer is used, which consists of 2 bidirectional LSTM layers of 128 units and 64 units respectively. It has the same 20% dropout rate for each recurrent block, 64-unit ReLU deep dense layer, and 24-unit linear output layer. This design guarantees that the only benefit to the baseline LSTM is the benefit of bidirectional temporal processing.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Hybrid CNN-BiLSTM-Attention Model

The proposed architecture directly addresses the future-work recommendations in the foundational research, which is a combination of convolutional feature extraction, bidirectional recurrence, and self-attention in a single architecture. There are three main phases involved with the network. The input sequence is first passed through two consecutive 1-D convolutional layers, one with the 64 filters and the other with 64 filters. It is a front-end that creates a local feature extractor and de-noising filter that can capture the patterns that occur in the short term (such as consumption peaks in the morning or evening load ramps). Second, the output of the convolved sequence is fed into a pair of stacked Bidirectional LSTM (128 and 64 units per direction) whose improved sequence modeling ability is utilized. Third, a multi-head self-attention block, surrounded by a residuals and layer normalization according to standard Transformer architecture, determines the most predictive relevant times in the look-back window of 168 historical hours to forecast the next hour.

The attended representation is then reduced to a single context vector using global average pooling and fed into a two-layer dense forecasting head with 128 and 64 ReLU units and dropout in between. Then, there is a linear output layer which gives the 24-step forecast horizon. The full architecture has 469,912 trainable parameters (approx. 1.79 MB), which is about 3.6 times the number of parameters in the baseline LSTM (as reported in Table 5).

Table 5. Proposed Hybrid CNN-BiLSTM-Attention model summary (469,912 trainable parameters).

Layer (type)	Output shape	Parameters
Input window	(None, 168, 15)	0
Conv1D	(None, 168, 64)	2,944
Conv1D_1	(None, 168, 64)	12,352
Dropout	(None, 168, 64)	0
Bidirectional	(None, 168, 256)	197,632
Dropout_1	(None, 168, 256)	0
Bidirectional_1	(None, 168, 128)	164,352
Multi-Head Attention	(None, 168, 128)	66,048
Add (residual)	(None, 168, 128)	0
Layer Normalization	(None, 168, 128)	256
Global Average Pooling	(None, 128)	0
Dense	(None, 128)	16,512
Dropout_3	(None, 128)	0
Dense_1	(None, 64)	8,256

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Layer (type)	Output shape	Parameters
Forecast output (Dense)	(None, 24)	1,560

Training Procedure

All three models have a uniform training setup to isolate the architectural performance from the training variations. The Adam algorithm with an initial learning rate of 0.001 is used to optimize. The objective function is based on Mean Squared Error and Mean Absolute Error is a secondary measure. Both models are trained using batch size of 64, and for a maximum of 200 epochs. To avoid overfitting and help convergence, training also uses early stopping with a patience of 15 epochs that resets the weights that have the smallest validation loss. A ReduceLROnPlateau callback is also triggered. Despite the six combinations of model configuration (including three different architectures and two different horizons), training automatically halts at a maximum of 17 epochs, a maximum of 41 epochs, proving that early stopping works as it should for all trials.

Evaluation Metrics

The evaluation of the performance is carried out with Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These metrics are calculated for the predictions both in the original MinMax scaled range and afterwards renormalized to megawatts. They are defined as:

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

In these expressions, y_i represents the actual load, $\hat{y}_i$ denotes the predicted load, and n stands for the total number of observations. Because RMSE squares differences before averaging, it penalizes larger errors more heavily than MAE, providing a balanced evaluation framework.

Results and Discussion

Training Behavior

The loss curves in Figure 3 are rescaled versions of training and validation loss for the three different architectures on both training and validation datasets, respectively, for both horizons. All models converge very rapidly within first 5 to 10 epochs. The LSTM and Bi-LSTM model start to deviate from each other after 5 or 6 epochs on the one-hour task, exhibiting some level of overfitting. The Hybrid model is different: while its validation loss is noisy, it doesn't steadily increase as the Hybrid model does. Extend the prediction out to 24 hours, and things change. The harder the multi-step task, the more stable and tighter the curves that all three models produce are, since they are unable to persist too long.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

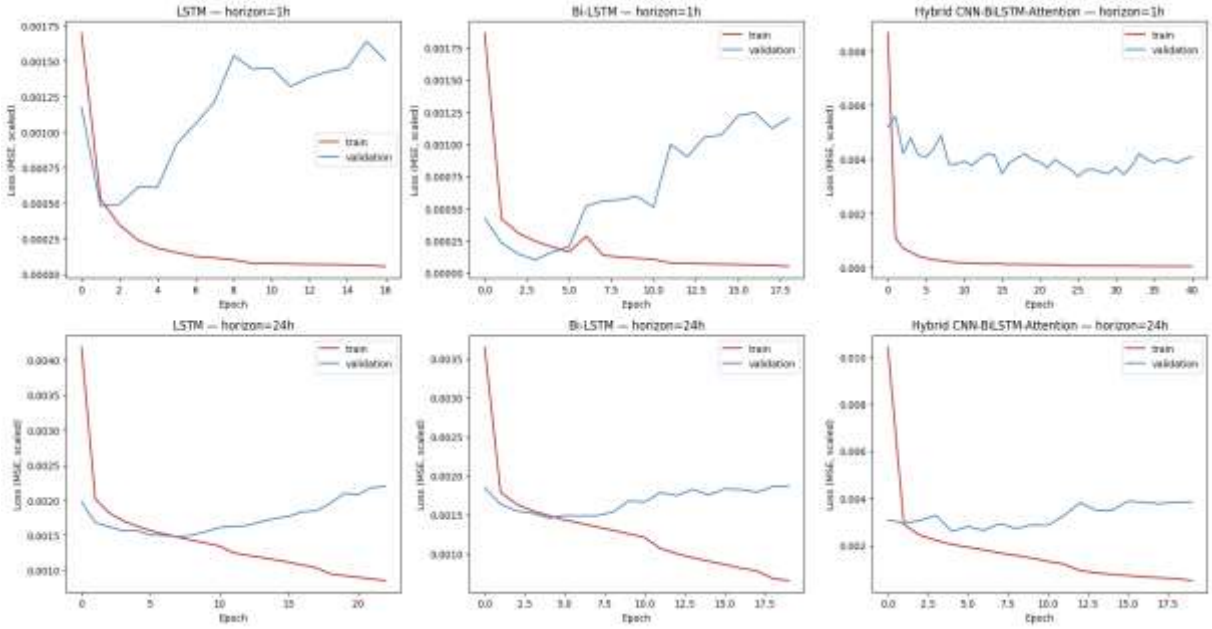


Figure 3. Training and validation loss (scaled MSE) for LSTM, Bi-LSTM, and Hybrid CNN-BiLSTM-Attention at the 1-hour (top row) and 24-hour (bottom row) forecasting horizons.

Overall Model Performance

The results of error metrics show that there are clear performance differences across the training and test partitions for both forecast horizons for the architectures. The Bi-LSTM gives the lowest error of 528.17 MW and 402.62 MW with h = 1 on the test set. The standard LSTM outperforms the Bi-LSTM with an RMSE of 2014.78 MW while the Bi-LSTM is obtaining an RMSE of 2033.93 MW at h = 24. The Hybrid CNN-BiLSTM-Attention model hits the highest test error at both horizons.

Table 6. Overall model performance summary (train and test, both horizons).

Model	H	Set	MAE (scaled)	RMSE (scaled)	MAE (MW)	RMSE (MW)
Bi-LSTM	1	Train	0.0079	0.0105	375.54	498.72
Bi-LSTM	**1**	**Test**	**0.0085**	**0.0111**	**402.62**	**528.17**
LSTM	1	Train	0.0165	0.0221	781.85	1050.09
LSTM	1	Test	0.0155	0.0209	736.76	990.72
Hybrid CNN-BiLSTM-Attn.	1	Train	0.0546	0.0619	2593.58	2936.59
Hybrid CNN-BiLSTM-Attn.	1	Test	0.0487	0.0560	2312.13	2658.80
Bi-LSTM	24	Train	0.0258	0.0356	1225.97	1687.79

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Model	H	Set	MAE (scaled)	RMSE (scaled)	MAE (MW)	RMSE (MW)
Bi-LSTM	24	Test	0.0314	0.0429	1488.30	2033.93
LSTM	24	Train	0.0251	0.0347	1189.99	1645.10
LSTM	**24**	**Test**	**0.0311**	**0.0424**	**1475.48**	**2014.78**
Hybrid CNN-BiLSTM-Attn.	24	Train	0.0383	0.0515	1818.95	2445.60
Hybrid CNN-BiLSTM-Attn.	24	Test	0.0385	0.0520	1829.71	2467.19

On aggregate accuracy alone, the two recurrent baselines outperform the Hybrid model; Sections 4.3 - 4.5 show that this aggregate picture is incomplete once cross-horizon behaviour is considered.

Architecture Comparison Within Each Horizon

Holding the forecast horizon fixed isolates which architecture is most accurate at a given task difficulty. Table 7 compares the three models at h = 1; Table 8 compares them at h = 24.

Table 7. Architecture comparison at horizon h = 1 (test set).

Model	MAE (scaled)	RMSE (scaled)	MAE (MW)	RMSE (MW)
Bi-LSTM	0.0085	0.0111	402.62	528.17
LSTM	0.0155	0.0209	736.76	990.72
Hybrid CNN-BiLSTM-Attn.	0.0487	0.0560	2312.13	2658.79

Bi-LSTM is the clear leader at h = 1, reducing RMSE by 46.7% relative to LSTM and by 80.1% relative to the Hybrid model. The Hybrid model's convolutional and attention components do not translate into an accuracy advantage at this short horizon under the current training configuration.

Table 8. Architecture comparison at horizon h = 24 (test set).

Model	MAE (scaled)	RMSE (scaled)	MAE (MW)	RMSE (MW)
LSTM	0.0311	0.0424	1475.48	2014.78
Bi-LSTM	0.0314	0.0429	1488.30	2033.93
Hybrid CNN-BiLSTM-Attn.	0.0385	0.0520	1829.71	2467.19

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Both LSTM and Bi-LSTM are almost the same at h = 24 (the difference in the RMSE value is less than 1 percent).The difference between the two models is less than 1 percent at h = 24 with the LSTM slightly outperforming the Bi-LSTM. The error corresponding to this horizon is again the highest with the Hybrid model, albeit with a lower relative error (22.5% higher RMSE than LSTM, compared to 80.1% higher RMSE than Bi-LSTM at h = 1) than in other cases across the horizon, as will be discussed in Section 4.5.

Prediction Visualization

The actual and predicted 24 hours ahead load are plotted on top of each other in Figure 4 for five consecutive test windows for each architecture. In all three models, the load series' overall daily and weekly cyclical pattern is followed. The prediction paths of LSTM and Bi-LSTM are closer to the peaks and troughs of the actual series, especially around the sharp peaks in the evenings, whereas the predictions of the Hybrid model appear smoother and generally under-predict the highest peaks, as reflected in the higher aggregate MAE and RMSE in Table 6.

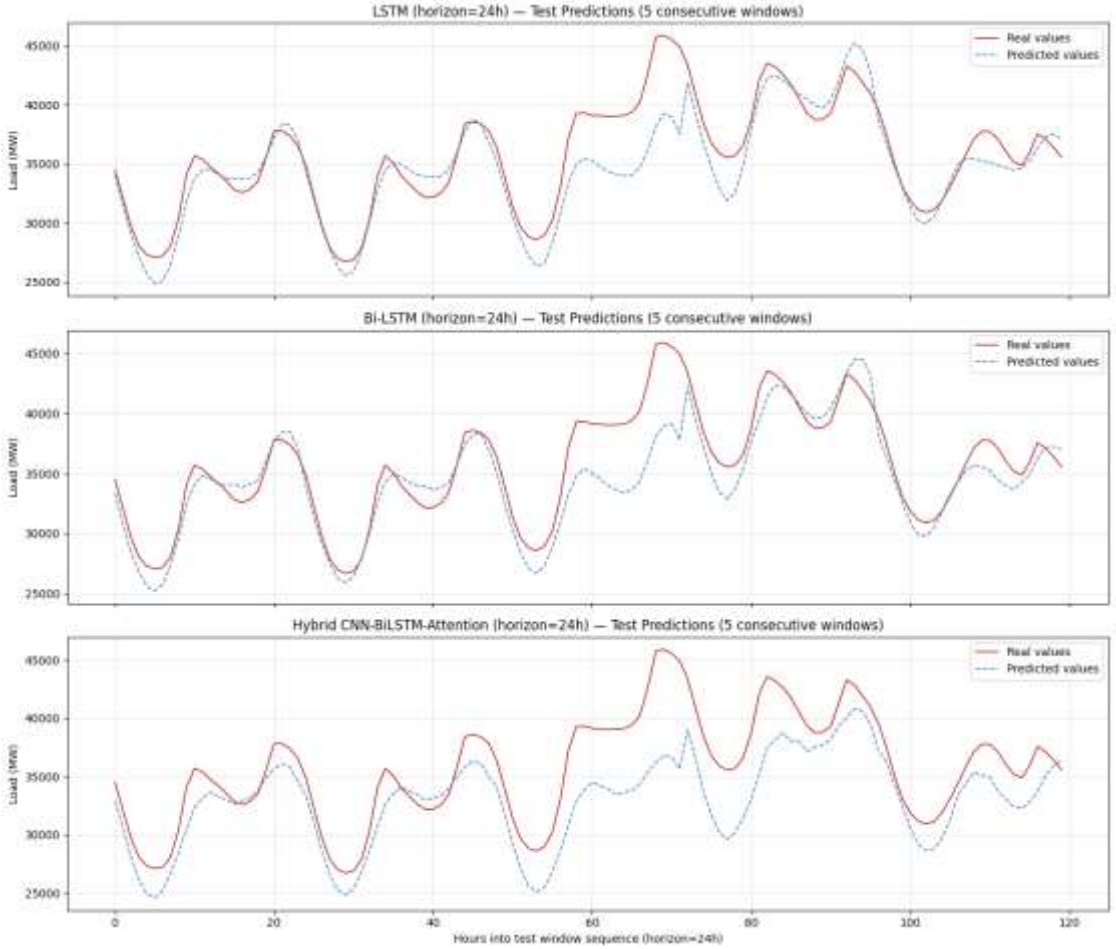


Figure 4. Actual versus predicted load for the 24-hour-ahead horizon, five consecutive test windows, for LSTM (top), Bi-LSTM (middle), and Hybrid CNN-BiLSTM-Attention (bottom).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

One-Step versus Multi-Step Forecasting Comparison

This comparison directly answers the base paper's proposed extension into multi-step forecasting: for each architecture, the h = 1 test error is compared with the h = 24 test error and expressed as a percentage change. Because a 24-hour-ahead forecast is intrinsically harder than a 1-hour-ahead forecast short-horizon persistence ("the next hour looks like this hour") is a strong predictor only at very short horizons absolute error is expected to increase for every architecture. The architecture whose error grows least is the one that relies most on genuine periodic structure (daily and weekly lag positions) rather than on short-horizon persistence.

Table 9. One-step (h = 1) versus multi-step (h = 24) test RMSE, by architecture.

Architecture	Test RMSE, h = 1 (MW)	Test RMSE, h = 24 (MW)	% change
LSTM	990.72	2014.78	+103.4%
Bi-LSTM	528.17	2033.93	+285.1%
Hybrid CNN-BiLSTM-Attn.	2658.79	2467.19	−7.2%

Both recurrent baselines show large error growth when the horizon extends from 1 to 24 hours: LSTM's test RMSE increases by 103.4% and Bi-LSTM's by 285.1%. The Hybrid CNN-BiLSTM-Attention model, in contrast, shows a small decrease of 7.2% in test RMSE (and a 20.9% decrease in test MAE) over the same change in horizon. This reflects a genuine trade-off between absolute accuracy and cross-horizon stability rather than the Hybrid model being unconditionally worse: at h = 1, Bi-LSTM's very low error is driven substantially by short-horizon persistence, a heuristic that stops being informative at h = 24 and so its error grows sharply; the Hybrid model's attention mechanism, by weighting daily and weekly lag structure directly (Section 4.6), is comparatively insensitive to this loss of persistence signal, so its error at the two horizons is far more similar. Figure 5 visualizes this pattern directly.

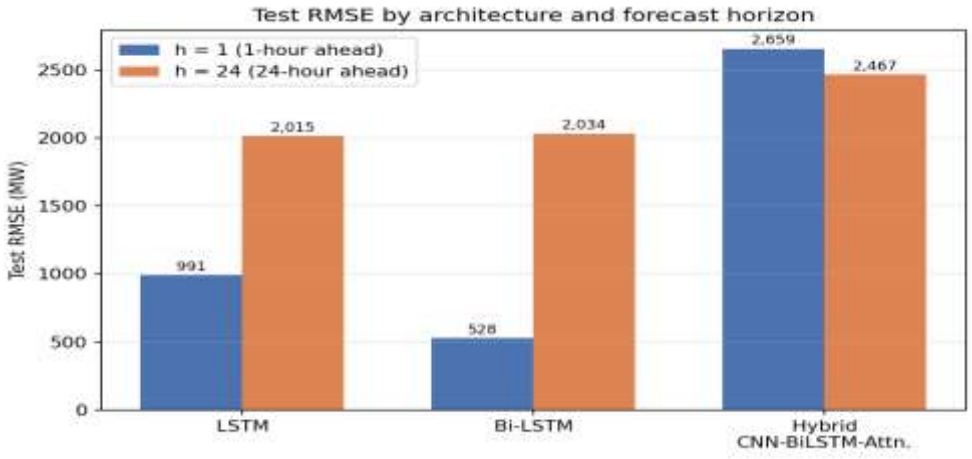


Figure 5. Test RMSE by architecture at the 1-hour and 24-hour forecasting horizons, illustrating the accuracy stability trade-off.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 1 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Interpretability: Attention Weight Analysis

The multi-head self-attention block in the Hybrid model offers a partial answer to the interpretability limitation noted in the base paper, since the weight the model assigns to each of the 168 input hours can be inspected directly rather than relying on an LSTM's opaque final hidden state. Figure 6 plots the average attention weight (across heads and query positions, for a representative test window) against hours relative to the forecast origin, at both h = 1 and h = 24, with reference lines marking the 24-hour and 168-hour (one-week) lag positions.

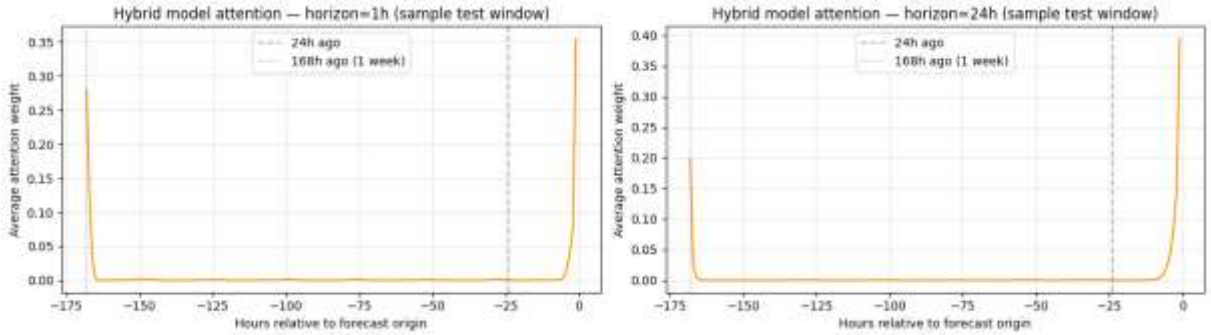


Figure 6. Average attention weight by lag position for the Hybrid model at h = 1 (left) and h = 24 (right).

The model focuses on the most recent hour of input for both horizons, with a relatively strong secondary peak at the 168-hour (one-week) lag, and relatively weak secondary peaks at the 24-hour and 48-hour lags. This shows that the information from the most recent observation is the key element in forecasts from the Hybrid model, and that this information holds the most importance in the forecast up to 24 hours ahead, which is consistent with the observed increased stability of the Hybrid model in longer lead times, as weekly-recurrence information is still useful up to 168 hours ahead, but daily-recurrence information would not be.

Discussion

Aggregate Accuracy versus Cross-Horizon Stability

The trend over Sections 4.2-4.5 is well defined, with an absolute accuracy trade-off in exchange for a cross-horizon stability. While the absolute errors of Bi-LSTM and LSTM are smallest at the optimal horizon, their errors increase rapidly (by 103-285%) when the horizon is increased from 1 to 24 hours. The Hybrid CNN-BiLSTM-Attention model achieves better absolute error values throughout, with the error value being almost constant, or even slightly decreasing, in the same horizon change. Both properties are not inherently 'better' – it is a case of which is more important depending on the application scenario.

Feature Engineering versus Architectural Complexity

The improved accuracy relative to what a purely univariate model would be expected to achieve appears attributable primarily to the engineered multivariate feature set, cyclical time encodings, lag features, and rolling statistics, rather than to increased architectural complexity. Under the present training configuration, the simplest recurrent architectures (LSTM, Bi-LSTM) outperform the far larger (3.6x the

Online ISSN 3007-3197
Print ISSN 3007-3189

parameters) Hybrid model on aggregate accuracy, suggesting that, once informative lag and rolling features are already present in the input, additional convolutional and attention machinery does not automatically translate into lower absolute error; its benefit here shows up instead in cross-horizon stability and interpretability.

Practical Implications

These findings have direct relevance to demand-supply planning in power systems facing chronic imbalance, such as Pakistan's grid. For applications that require the most accurate possible short-term (hour-ahead) forecast, for example, real-time dispatch or frequency regulation, Bi-LSTM is the most suitable of the three models evaluated here. For day-ahead (24-hour) generation scheduling, LSTM and Bi-LSTM perform almost identically, with LSTM marginally ahead. Where forecasts must be produced reliably across a range of horizons, or where operators need insight into which historical periods are driving a given forecast (for example, to validate a forecast against known demand drivers before committing generation resources), the Hybrid CNN-BiLSTM-Attention model's stability and attention-based interpretability offer a meaningful, if less numerically accurate, alternative.

Limitations

This study relies exclusively on the PJM East dataset, meaning findings require validation before generalizing to other power grids like Pakistan's. Because the analysis uses hourly load series rather than the Enefit prosumer data from the base paper, absolute error values differ, making this work an independent test of the architecture family rather than a direct replication. The feature set is limited to historical load and timestamp data, as other data including economic indicators and public holidays are not included in the data set, nor are other external factors such as weather. The architecture of Hybrid CNN-BiLSTM-Attention model is very simple, with only one configuration of hyper-parameters, which implies that there is a potential opportunity that systematic tuning could fill the gap in accuracy between the recurrent baselines. The comparison only includes three architectures, and not recent models such as Transformer forecasters. Last but not least, the evaluation is still completely simulation-based, without any real-world operational problems such as concept drift, data latency, or periodic model retraining.

Conclusion

This study evaluated LSTM, Bi-LSTM, and a proposed Hybrid CNN-BiLSTM-Attention model for electricity load forecasting on the PJM East hourly load series, using a fifteen-column multivariate feature set and two forecasting horizons (1-hour-ahead and 24-hour-ahead), directly extending the univariate, single-horizon comparison of LSTM and Bi-LSTM reported in the base paper. Bi-LSTM achieved the lowest error for 1-hour-ahead forecasting (test RMSE = 528.17 MW), while LSTM was marginally the most accurate model for 24-hour-ahead forecasting (test RMSE = 2014.78 MW, versus 2033.93 MW for Bi-LSTM). The Hybrid CNN-BiLSTM-Attention model recorded the highest absolute error at both horizons but showed markedly better cross-horizon stability, with its test RMSE changing by only 7.2% between h = 1 and h = 24, compared with +103.4% for LSTM and +285.1% for

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Bi-LSTM, and it provided a transparent, attention-based account of which historical hours' drive each forecast, most weight on the most recent hour, with a secondary peak at the one-week lag. In general, results suggest that the multivariate input features thoughtfully designed had a greater impact on forecasting accuracy in this study than the added complexity of the architecture, and the attention-based hybrid architecture incurred a trade-off between absolute accuracy and stability and interpretability across forecasting horizons. The findings can be used in the design of practical load-forecasting systems, especially in power systems with a high demand-supply shortage.

Future Work

Targeted extensions are suggested for future work building on the above findings. External auxiliary information such as weather measurements, holiday calendars and economic indicators can be incorporated in the research to reflect the wider influences on demand. A potential improvement is to fine-tune hyper-parameters of the Hybrid CNN-BiLSTM-Attention model systematically, which can lower the absolute error and obtain the stability benefits. These architectures are validated on datasets from emerging grids like Pakistan, demonstrating their suitability to address distinct challenges in Pakistan, such as heavy load shedding and seasonal change. The evaluation horizons of 24 hours are extended to 48 hours or one week, which make it easier to understand the limits of long term stability. Training an independently windowed one-step-ahead model, not the multi-step output makes the implementation more closely match the base paper. Short-term recurrent precision and cross-horizon stability is achieved through exploring Transformer designs and hybrid ensembles.Short-term recurrent precision is achieved while exploring Transformer designs and hybrid ensembles, along with cross-horizon stability.

References

1. Lodhi, R.N. and R. Malik, *Impact of electricity shortage on daily routines: A case study of Pakistan.* Energy & environment, 2013. **24**(5): p. 701-709.
2. Dudek, G., P. Piotrowski, and D. Baczyński, *Forecasting in modern power systems: challenges, techniques, and emerging trends*. 2025, MDPI. p. 3589.
3. Misiurek, K., T. Olkuski, and J. Zyśk, *Review of methods and models for forecasting electricity consumption.* Energies, 2025. **18**(15): p. 4032.
4. Cui, D., et al., *Enhancing short-term electricity forecasting with advanced machine learning techniques.* Journal of Electrical Engineering & Technology, 2026. **21**(1): p. 147-187.
5. Alizadegan, H., et al., *Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction.* Energy Exploration & Exploitation, 2025. **43**(1): p. 281-301.
6. Park, J. and J. Jeong, *An autoscaling system based on predicting the demand for resources and responding to failure in forecasting.* Sensors, 2023. **23**(23): p. 9436.
7. Shapi, M.K.M., N.A. Ramli, and L.J. Awalin, *Energy consumption prediction by using machine learning for smart building: Case study in Malaysia.*

Online ISSN 3007-3197
Print ISSN 3007-3189

Developments in the Built Environment, 2021. **5**: p. 100037.

8. Khalil, M., et al., *Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review.* Engineering Applications of Artificial Intelligence, 2022. **115**: p. 105287.
9. Momani, M.A., et al., *Comparative analysis of ARIMA and SARIMA models in electrical load forecasting: Insights for long and short-term projections.* International Journal of Electrical Engineering and Computer Science, 2025. **7**: p. 83-89.
10. Alzoubi, A., *Machine learning for intelligent energy consumption in smart homes.* International Journal of Computations, Information and Manufacturing (IJCIM), 2022. **2**(1).
11. Wu, M.-P. and F. Wu, *Predicting residential electricity consumption using CNN-BiLSTM-SA neural networks.* IEEE Access, 2024. **12**: p. 71555-71565.
12. Bilgili, M. and E. Pinar, *Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye.* Energy, 2023. **284**: p. 128575.
13. Ravivarma, K., et al., *A comparative study of machine learning and deep learning models for renewable energy and load demand forecasting in smart microgrids.* Discover Sustainability, 2026. **7**(1): p. 518.