Improving the prediction of hourly electricity consumption using deep learning models

Antriksh Dhand, Tommy Nguyen, Srijan Chaudhary, Pranav Lokhande, Saptarshi Roy,
Jainil Chauhan, Josiah Badham and Aviral Singhal

ARTICLE INFO

Keywords:
building energy consumption
energy prediction
Conv-LSTM
LGBM
deep learning
LSTM
BDG2

ABSTRACT

Buildings account for nearly 40% of global energy consumption, making accurate load forecasting critical for energy efficiency and decarbonisation efforts. This paper presents a comparative evaluation of four machine learning models – Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), Convolutional LSTM (ConvLSTM), and Light Gradient-Boosting Machine (LightGBM) – for 24-hour ahead building electricity forecasting. Using the Building Data Genome 2 dataset comprising 20 diverse commercial buildings, we assess model performance across R^2, RMSE, MAE, and computational cost. LightGBM achieved the highest overall accuracy ($R^2 = 0.627$, RMSE $=$ 34.65 kWh), outperforming deep learning models in 80% of buildings, while BiLSTM excelled on highly irregular or low-demand profiles. All models substantially improved upon naive baselines, explaining over 60% of consumption variance. A composite score combining accuracy, error, and training time identified LightGBM as the most balanced architecture for operational deployment. These results highlight that gradient boosting offers a robust, scalable alternative to deep sequence models, providing actionable insights for intelligent building management and energy forecasting applications.

1. Introduction

Buildings account for roughly 36-40% of global energy use and associated carbon emissions, making them a central focus of decarbonisation efforts [2, 1, 15]. Accurate short-term prediction of building electricity consumption supports load-balancing, operational optimisation, and integration of renewable energy that can be used to reduce a building's environmental footprint.

Traditional physics-based (*white-box*) simulators such as *EnergyPlus* and *TRNSYS* provide interpretability, but require detailed inputs and heavy computation. In contrast, data-driven (*black-box*) models learn directly from historical data, allowing scalable forecasting in diverse buildings [8, 16]. Recent advances in deep learning – particularly CNNs, LSTMs, and their hybrids – have significantly improved accuracy, while ensemble methods such as LightGBM offer competitive results with lower computational cost [7].

This paper provides a unified and leak-free comparison of CNN, BiLSTM, ConvLSTM, and LightGBM models on the BDG2 dataset [12]. All models are trained with identical preprocessing, features, and evaluation metrics (RMSE, R^2, CV-RMSE). The study benchmarks predictive accuracy, computational efficiency, and robustness in 20 buildings, clarifying trade-offs between deep learning and gradient-boosting approaches for practical energy forecasting.

2. Background

Global energy consumption has increased sharply in recent decades, intensifying concerns about climate change, dependence on fossil fuels, and sustainable resource management. Among all end-use sectors, buildings account for an especially large and growing share of this demand, contributing roughly 36-40% of global energy consumption and carbon emissions [2, 1, 15]. This makes the building sector both a major challenge and a clear opportunity in the pursuit of energy conservation and decarbonisation. Thus, improving the energy efficiency of buildings has become a central pillar in global sustainability and carbon neutrality initiatives.

Accurate prediction of building energy consumption is one of the most effective strategies to support these efforts. Energy forecasting models enable building operators, urban planners, and policymakers to anticipate energy demand, optimise building operation, and design targeted energy-saving strategies [15, 2]. By revealing the underlying consumption patterns of buildings, such models allow for demand-side management, load balancing, and informed decision-making for energy-efficient design. Energy prediction also assists smart-grid integration and renewable energy scheduling by aligning supply with anticipated demand, thereby reducing operational costs and improving system reliability.

The energy consumption of a building is influenced by a complex interplay of factors, including HVAC efficiency, occupant behaviour, equipment load, and external weather conditions [1]. This multivariate dependency makes accurate energy prediction a challenging task, requiring models capable of capturing both temporal dynamics and nonlinear relationships among variables. Traditional approaches (often referred to as *physical* or *white-box* models) rely on detailed information about the geometry, materials, and systems of a building to simulate thermal behaviour using tools such as *EnergyPlus* or *TRNSYS* [2, 8]. Although physically grounded models offer interpretability and engineering transparency, they are computationally intensive, require extensive input data, and often struggle to reflect real-time occupant and environmental variability [14].

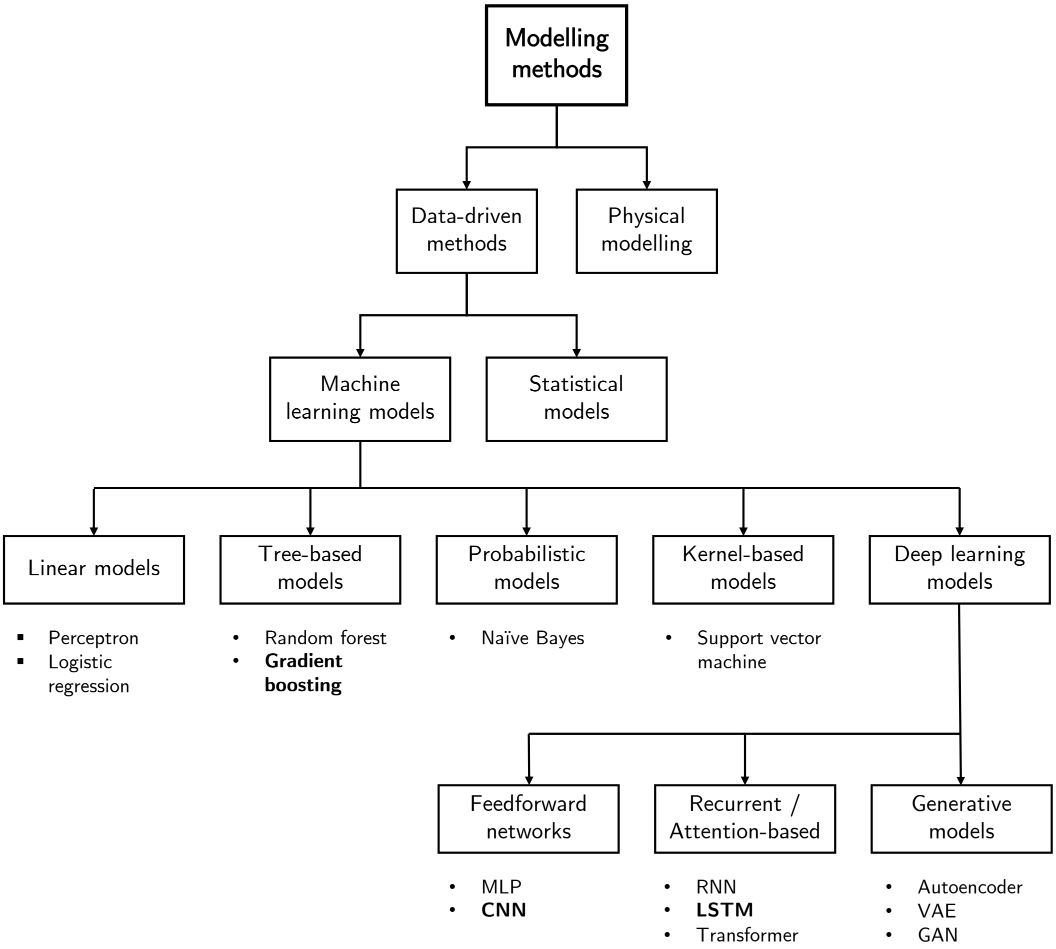

Figure 1: Various methods used for electricity consumption prediction. Methods used in this paper are in bold.

To address these limitations, researchers have increasingly turned to *data-driven* (or *black-box*) approaches, which infer relationships directly from empirical data rather than relying on explicit physical formulations [8]. Whereas physical models depend on predefined parameters – such as design data, thermal properties, and HVAC configurations – data-driven models learn patterns autonomously from historical observations of energy use and contextual factors [16]. These methodologies can be broadly categorised into two groups: (a) statistical models and (b) machine learning models. Although statistical approaches remain more interpretable, recent studies have increasingly favoured machine learning techniques due to their superior predictive accuracy. Within this domain, deep learning methods have become particularly effective in predicting electricity consumption. A summary of the major modelling paradigms is presented in Figure 1.

The early applications of machine learning in building energy forecasting established the value of data-driven methods. Yang et al. (2005) introduced adaptive online neural predictors capable of tracking operational drift [24], while Dong et al. (2005) demonstrated that Support Vector Machines outperform traditional statistical models in tropical climates [4]. Li et al. (2009) expanded this comparison to multiple variants of Artificial Neural Network (ANN), standardising short-horizon forecasting practices [9]. Wong et al. (2010) subsequently showed that ANNs can emulate *EnergyPlus* simulations with high fidelity [23], and Zhao & Magoulès (2012) later synthesised these developments, advocating for rigorous evaluation and hybrid (grey-box) approaches [27].

Since then, architectural advances (particularly in recurrent and convolutional networks) have enabled models such as LSTMs, BiLSTMs, and CNNs to learn long-range dependencies and cyclical consumption motifs. Hybrid architectures such as ConvLSTM and attention-augmented variants further enhanced predictive capacity by integrating local feature extraction with temporal context learning. Parallel to these approaches, ensemble methods such as Gradient-Boosted Trees have gained traction for their robustness, interpretability, and efficiency, offering competitive performance on structured tabular data typical of building energy records.

However, despite these advances, there is still a persistent methodological gap. Many studies focus narrowly on improving model architecture without ensuring fair comparison between algorithms, consistent preprocessing, or reproducible evaluation frameworks. As Zhao & Magoulès (2012) originally cautioned, progress in model sophistication has often outpaced methodological rigour. Recent reviews reaffirm this concern, emphasising the lack of standardised pipelines that enable direct "apples-to-apples" benchmarking across machine learning and deep learning methods [15].

This study seeks to address this gap by conducting a systematic, reproducible comparison of four representative architectures – CNN, BiLSTM, ConvLSTM, and LightGBM – on the publicly available Building Data Genome 2 dataset. By keeping data processing, feature engineering, and evaluation procedures constant, this work aims to clarify the relative strengths, weaknesses, and practical deployment potential of both deep sequence models and gradient-boosted ensembles for short-term building energy forecasting.

3. Methodology

3.1. Data

Artificial intelligence techniques attempt to mimic the human capacity for inductive learning; that is, machine learning is based on learning by example [19]. The more data provided to these architectures, the better their ability to identify underlying patterns and generalise to unseen cases. In contrast, limited training data constrain the accuracy and robustness of the model. This property is particularly pronounced in deep learning architectures that require an even greater number of training data points to achieve the desired accuracy and generalisation performance [20]. Consequently, the availability of large, high-quality datasets is essential for data-driven energy consumption prediction. Several public datasets have been released to support this effort, including the UCI Energy Efficiency Dataset [22] and the Pecan Street Dataset [17].

The dataset used in this study is a subset of Building Data Genome 2 (BDG2), a large-scale repository of electricity consumption data collected from 1,636 non-residential buildings in 19 sites in North America and Europe [12]. Each site, denoted by an anonymised "animal" codename (e.g. Panther, Robin, Fox), represents a university campus or group of buildings (Table 1). Within these sites, one or more electrical meters were installed per building, capturing hourly readings over two full years (2016-2017). BDG2 also includes metadata describing each building's floor area, primary use type, geographic location, and time zone, along with hourly weather data from the nearest meteorological stations. In total, the dataset contains more than 53 million hourly electricity measurements from 3,053 meters, allowing for analysis across diverse types of buildings and climate zones.

Originally curated as part of the Building Data Genome Project, the dataset was used in the ASHRAE Great Energy Predictor III competition on Kaggle, designed to benchmark machine learning approaches for long-term building energy prediction. Given its extensive scale, diversity of building types, comprehensive metadata and open accessibility, BDG2 has since become a benchmark dataset for developing data-driven building energy prediction models.

3.1.1. Data processing

The data preprocessing methodology employed in this study was designed to closely replicate the approach of Liang et al. [10] to ensure methodological rigour and allow for a point of comparison with previous research.

1. The five most common building types were retained to avoid sparsely populated categories: *Education* (education), *Office* (office), *Entertainment/public assembly* (assembly), *Lodging/residential* (lodging), and *Public services* (public).
2. The weather dataset was reduced to the three variables with the fewest missing values: wind speed, dew temperature, and air temperature.
3. Buildings exhibiting more than 10% missing hourly electrical readings were excluded.
4. Buildings with a mean consumption below 20 kWh were excluded.
5. Key temporal features, namely month, day of week and hour, were extracted to capture cyclical patterns in energy use.

After this cleaning process, 398 educational buildings, 201 office buildings, 100 assembly buildings, 91 lodging buildings, and 106 public buildings remained. 20 buildings were then selected at random to form the final dataset (Table 2).

3.1.2. Random forest imputation

After data cleaning, a small proportion of missing values remained in several variables: electricity consumption (2.50%), wind speed (0.74%), dew temperature (0.62%), and air temperature (0.59%). An iterative random forest (RF) procedure was implemented to address this. The procedure processed each variable sequentially using all other variables as input features to train individual RF models for each variable with missing data. This approach follows Liang et al.'s methodology and provides a sophisticated alternative to simple mean or median imputation. It is particularly effective because it can capture non-linear relationships between variables and handle the complex interactions between weather conditions, temporal patterns, and building characteristics that influence electricity consumption.

For each target variable, missing values in the predictor variables were initially replaced with their mean values to create a complete training set. A random forest regressor was then trained using 100 estimators, a maximum depth of 10, a minimum samples split of 5, and a minimum samples leaf of 2, with parallel processing enabled for computational efficiency. This was subsequently repeated for each predictor variable containing missing values. The iterative nature of this process ensures that as each variable is imputed, subsequent variables benefit from the improved data quality

Table 1
Overview of BDG2 dataset sites, their locations, and number of buildings. [12]

Site	Actual site name	Location	Buildings
Panther	Univ. of Central Florida (UCF)	Orlando, FL, USA	136
Robin	Univ. College London (UCL)	London, UK	52
Fox	Arizona State Univ. (ASU)	Tempe, AZ, USA	137
Rat	Washington DC - City Buildings	Washington, DC, USA	305
Bear	Univ. of California - Berkeley	Berkeley, CA, USA	92
Lamb	Cardiff - City Buildings	Cardiff, UK	147
Eagle	Anonymous	N/A	47
Moose	Ottawa - City Buildings	Ottawa, ON, Canada	15
Gator	Anonymous	N/A	74
Bull	Univ. of Texas - Austin	Austin, TX, USA	124
Bobcat	Anonymous	N/A	36
Crow	Carleton Univ.	Ottawa, ON, Canada	5
Wolf	Univ. College Dublin (UCD)	Dublin, Ireland	36
Hog	Anonymous	N/A	163
Peacock	Princeton University	Princeton, NJ, USA	106
Cockatoo	Cornell University	Ithaca, NY, USA	124
Shrew	UK Parliament	London, UK	9
Swan	Anonymous	N/A	21
Mouse	Ormond Street Hospital	London, UK	7

Table 2
Summary of selected buildings grouped by category and site.

Category	Site	Building
Education	Bear	Bear_education_Chana
	Fox	Fox_education_Heriberto
	Lamb	Lamb_education_Robin
	Peacock	Peacock_education_Robbie
	Rat	Rat_education_Nellie
Assembly	Lamb	Lamb_assembly_Librada
	Rat	Rat_assembly_Kimberley
Lodging	Hog	Hog_lodging_Shanti
	Peacock	Peacock_lodging_Sergio
		Peacock_lodging_Mathew
	Robin	Robin_lodging_Elmer
Office	Hog	Hog_office_Lizzie
		Hog_office_Richelle
		Hog_office_Joey
		Hog_office_Elsy
	Peacock	Peacock_office_Annie
	Robin	Robin_office_Shirlene
Public	Fox	Fox_public_Rhonda
	Hog	Hog_public_Octavia
	Rat	Rat_public_Margart

of previously imputed variables. This creates a cascading effect where the accuracy of imputation improves with each iteration. Finally, the trained model was subsequently used to predict and replace missing values in the target variable.

The imputed values maintain statistical consistency with the original data, with mean values remaining virtually unchanged and standard deviations showing minimal variation, indicating that the RF predictions preserve the underlying data distribution characteristics (Table 3).

3.1.3. *Final dataset characteristics*

After completing the RF imputation procedure, the final dataset contained 877,200 hourly observations across 11 variables with zero missing values. The final dataset includes the hourly timestamp, building identifiers (site_id,

Table 3
Variable means and standard deviations before and after random forest imputation.

Variable	Original Mean	Imputed Mean	Δ Mean	Original Std	Imputed Std	Δ Std
Electricity consumption	256.38	256.95	+0.57	422.05	417.08	−4.97
Wind speed	3.55	3.55	0.00	2.32	2.31	−0.01
Dew temperature	7.14	7.16	+0.02	10.29	10.27	−0.02
Air temperature	14.09	14.09	0.00	10.70	10.67	−0.03

building_id, building_type), weather variables (wind_speed, dew_temp, air_temp), temporal features (month, dayofweek, hour), and the target electricity consumption variable.

3.2. Classifier architectures

The four architectures we explore in this paper – CNNs, BiLSTMs, ConvLSTMs, and LightGBM – span complementary forecasting paradigms. They capture local, sequential, hybrid, and non-neural modelling strategies, respectively, allowing a balanced comparison across neural and ensemble methods.

All model architectures presented in the following sections follow the general input-output configuration and training procedure described in Section 3.3.

3.2.1. Convolutional Neural Network

Convolutional Neural Networks (CNNs) have shown strong performance in temporal pattern recognition by automatically learning hierarchical features and capturing local dependencies through parallel computation [3].

Our CNN network comprises three sequential 1D convolutional layers with 48, 96, and 128 filters and kernel sizes of 5, 3, and 3, respectively. These layers progressively extract short-, mid-, and long-term temporal features such as hour-to-hour changes and daily cycles. Each convolutional layer uses zero-padding to preserve temporal resolution, ReLU activation, and dropout regularisation ($p = 0.2$) to prevent overfitting.

Global average pooling (GAP) then aggregates features across the temporal dimension, producing a fixed-length 128-dimensional vector that summarises the learned representations. GAP encourages the network to learn temporally-invariant features that are useful regardless of their position in the input sequence, improving robustness to phase shifts in consumption patterns.

Finally, a fully connected forecasting head with dense layers of 128, 64, and 24 units maps these features to the 24-hour prediction horizon. The first two layers use ReLU activation and dropout ($p = 0.2$), and the output layer is linear. This progressive dimension reduction compresses learned temporal representations into the target 24-hour forecast horizon.

More formally, let $\mathbf{X} \in \mathbb{R}^{N\times24\times7}$ denote the input batch. The convolutional layers progressively extract features:

$$\mathbf{H}^{(l)} = \text{Dropout}(\text{ReLU}(\text{Conv1D}(\mathbf{H}^{(l-1)}))) \tag{1}$$

for $l \in \{1, 2, 3\}$ with $\mathbf{H}^{(0)} = \mathbf{X}$, output channels $\{48, 96, 128\}$, and kernel sizes $\{5, 3, 3\}$ respectively. Global average pooling compresses the temporal dimension:

$$\mathbf{z} = \frac{1}{24}\sum_{t=1}^{24} \mathbf{H}^{(3)}[:, t, :] \in \mathbb{R}^{N\times128} \tag{2}$$

The fully connected forecasting head processes the pooled features through three dense layers with output dimensions 128, 64, and 24 respectively, producing the final prediction $\hat{\mathbf{y}} \in \mathbb{R}^{N\times24}$.

3.2.2. Bidirectional Long Short-Term Memory

The BiLSTM extends the conventional LSTM by processing input sequences in both forward and backward directions, allowing the model to capture temporal dependencies from past and future contexts within the 24-hour observation window [5]. This property is particularly beneficial for building energy forecasting, where consumption at a given hour may depend on both preceding and anticipated daily patterns.

Our architecture comprises a single bidirectional LSTM layer with 128 hidden units per direction (256 total). Each LSTM cell employs standard gating mechanisms to regulate information flow through the hidden and cell states, enabling selective retention of long-term dependencies. For forecasting, the final forward and backward hidden states are concatenated, forming a 256-dimensional representation that summarises the entire 24-hour input sequence.

This representation is passed through a three-layer fully connected forecasting head. The first two dense layers (64 units, ReLU activation) progressively compress the temporal representation, while the final linear layer outputs 24 real-valued predictions corresponding to the next 24 hours of electricity consumption.

If $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_{24}] \in \mathbb{R}^{N\times24\times7}$ denotes the input batch, then the bidirectional LSTM processes the sequence:

$$\overrightarrow{\mathbf{h}}_t, \overrightarrow{\mathbf{c}}_t = \text{LSTM}_{\text{fwd}}(\mathbf{x}_t, \overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{c}}_{t-1}) \tag{3}$$

$$\overleftarrow{\mathbf{h}}_t, \overleftarrow{\mathbf{c}}_t = \text{LSTM}_{\text{bwd}}(\mathbf{x}_t, \overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{c}}_{t+1}) \tag{4}$$

for $t \in \{1, \dots, 24\}$, where $\overrightarrow{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t \in \mathbb{R}^{128}$. The final hidden states are concatenated:

$$\mathbf{h}_{\text{final}} = [\overrightarrow{\mathbf{h}}_{24}; \overleftarrow{\mathbf{h}}_1] \in \mathbb{R}^{N\times256} \tag{5}$$

The fully connected forecasting head transforms this representation through successive layers:

$$\mathbf{f}_1 = \text{ReLU}(\mathbf{W}_1\mathbf{h}_{\text{final}} + \mathbf{b}_1), \qquad \mathbf{f}_1 \in \mathbb{R}^{N\times64} \tag{6}$$

$$\mathbf{f}_2 = \mathrm{ReLU}(\mathbf{W}_2\mathbf{f}_1 + \mathbf{b}_2), \qquad \mathbf{f}_2 \in \mathbb{R}^{N\times 64} \tag{7}$$

$$\hat{\mathbf{y}} = \mathbf{W}_3\mathbf{f}_2 + \mathbf{b}_3, \qquad \hat{\mathbf{y}} \in \mathbb{R}^{N\times 24} \tag{8}$$

where $\hat{\mathbf{y}}$ represents the predicted 24-hour consumption sequence.

3.2.3. Convolutional–Long Short-Term Memory

The Convolutional–LSTM (ConvLSTM) model combines convolutional and recurrent layers to capture both short-term and long-range temporal dependencies. This architectural design is expected to provide superior feature representations compared to either CNNs or LSTMs alone, potentially capturing both fine-grained hourly fluctuations and broader daily consumption trends.

The architecture begins with a single 1D convolutional layer (16 filters, kernel size $k = 3$) that serves as a lightweight feature preprocessor, transforming the original seven input features into 16 learned feature maps that emphasise local temporal patterns. This convolutional preprocessing is expected to create more informative input representations for the LSTM by highlighting relevant local patterns and reducing noise in the raw input features.

Formally, let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_{24}] \in \mathbb{R}^{N\times 24\times 7}$ denote the input batch. The convolutional layer produces

$$\mathbf{C} = \mathrm{ReLU}(\mathrm{Conv1D}_{16,k=3}(\mathbf{X})) \in \mathbb{R}^{N\times 24\times 16}. \tag{9}$$

The transformed features are then passed to a two-layer bidirectional LSTM with 128 hidden units per direction, producing 256-dimensional representations:

$$\overrightarrow{\mathbf{h}}_t^{(l)}, \overleftarrow{\mathbf{h}}_t^{(l)} = \mathrm{BiLSTM}^{(l)}(\mathbf{c}_t, \overrightarrow{\mathbf{h}}_{t-1}^{(l)}, \overleftarrow{\mathbf{h}}_{t+1}^{(l)}) \tag{10}$$

for layers $l \in \{1, 2\}$ and timesteps $t \in \{1, \dots, 24\}$, where $\overrightarrow{\mathbf{h}}_t^{(l)}, \overleftarrow{\mathbf{h}}_t^{(l)} \in \mathbb{R}^{128}$. This structure enables hierarchical learning of both short- and long-range dependencies while incorporating context from past and future timesteps. The final hidden states are concatenated:

$$\mathbf{h}_{\text{final}} = [\overrightarrow{\mathbf{h}}_{24}^{(2)}; \overleftarrow{\mathbf{h}}_1^{(2)}] \in \mathbb{R}^{N\times 256} \tag{11}$$

and passed through two ReLU-activated dense layers (64 units each) and a final linear layer to generate the 24-hour forecast.

3.2.4. Light Gradient-Boosting Machine

The Light Gradient-Boosting Machine (LightGBM) model serves as the non-neural baseline for short- and medium-term electricity consumption forecasting. LightGBM is a gradient-boosted decision tree algorithm that grows trees leaf-wise to improve accuracy and training efficiency relative to traditional boosting methods.

The feature set comprised of both temporal and weather-driven predictors. Lag features of 1-168 hours captured short-term, diurnal, and weekly dependencies, while rolling statistics (mean and standard deviation) over windows of 3-168 hours characterised local trends and variability. Weather variables (air temperature, dew point, wind speed) were included with 24-hour and 168-hour lags and corresponding rolling aggregates. Categorical attributes such as building id, building type, month, dayofweek, and hour were encoded using LightGBM's native categorical handling.

The final model used 31 leaves, learning rate = 0.05, maximum depth = 8, feature fraction = 0.9, bagging fraction = 0.9, L2 regularisation = 2.0, and a minimum of 300 samples per leaf.

3.3. Training procedure

This section outlines the data preparation, model training configuration, and evaluation procedures used to ensure consistent and comparable results across all architectures.

3.3.1. Input data

All models accept input sequences of 24 hourly timesteps, with seven features per timestep: electricity consumption, wind speed, dew point temperature, air temperature, month, day of week, and hour. This 24-hour lookback window captures one complete diurnal cycle, which is essential for learning daily consumption patterns. Each model outputs a continuous sequence of 24 hourly consumption forecasts corresponding to the next day.

All features were normalised using the `RobustScaler` implementation from `scikit-learn` [18], which scales features based on the interquartile range. This scaling method ensures that the feature distributions remain robust to occasional extreme values often caused by sensor errors or irregular operational conditions.

3.3.2. Training configuration

Hyperparameters were selected empirically through iterative testing and literature-based reference tuning to balance accuracy and computational efficiency. The final training configuration is summarised in Table 4.

Table 4
Final training parameters for all models.

Parameter	Final value
GPU batch size	64
Optimiser	Adam
Loss function	Mean Squared Error (MSE)
Learning rate	0.001
Validation approach	Leave-1-out 5-fold CV
Evaluation metrics	RMSE, MAE, R^2, CV-RMSE

All neural network models (CNN, BiLSTM, ConvLSTM) were trained using the Adam optimiser with learning rate $\alpha = 0.001$, minimising mean squared error loss over 8 epochs with batch size 128. The relatively modest epoch count was selected to prevent overfitting given the limited training sequences available per building (approximately 8,711 samples after sequence creation from 2016 data).

All models were implemented using PyTorch and trained on Google Colab using the T4 GPU to ensure consistency in the experiment.

3.4. Evaluation techniques

The training strategy followed a per-building approach: an independent model was trained for each building to capture unique consumption characteristics and operational behaviours.

3.4.1. *Train-test split*

For all architectures, the 2016 calendar year was used for training and validation, while 2017 was held as the test set to maintain temporal integrity (174,260 test samples).

It is common practice to divide the input data into k folds for cross-validation, where accuracy and loss are averaged across all k folds to provide a reliable estimate of the model's performance. Here, 5-fold cross-validation was performed on the 2016 training data. Each fold was trained independent of random initialisation, and the model that achieved the lowest validation mean squared error was selected for evaluation in the 2017 test set. This cross-validation strategy provides more reliable and generalisable performance estimates compared to single train-validation splits.

3.4.2. *Metrics*

Performance was evaluated using the Root Mean Square Error (RMSE, Equation 12), Mean Absolute Error (MAE, Equation 13), Coefficient of Determination (R^2, Equation 14), and Coefficient of Variation of RMSE (CV-RMSE, Equation 15).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

$$\text{CV-RMSE} = \frac{\mathrm{RMSE}}{\bar{y}} \times 100\% \tag{15}$$

where y_i and $\hat{y}_i$ denote the observed and predicted values for sample i, respectively, $\bar{y}$ is the mean of the observed values and n is the number of samples.

These were benchmarked against a persistence-24h baseline model that assumes the next day's consumption equals the current day's.

4. Results

4.1. Overall model performance

Table 5 presents the overall forecasting performance of each model averaged in all 20 buildings. LightGBM achieved the highest predictive accuracy, with a mean R^2 of 0.627 and substantially lower error metrics (RMSE = 34.65 kWh; MAE = 20.56 kWh) than the deep learning models. In comparison, the BiLSTM and ConvLSTM recurrent networks obtained R^2 values of 0.572 and 0.549, respectively. The CNN lagged behind, delivering the lowest mean R^2 (0.493) and the highest errors (RMSE = 43.79 kWh; MAE = 28.89 kWh). All models produced positive and moderate R^2 scores, which explained approximately 50–63% of the variance in hourly energy consumption. LGBM's performance was not only the highest on average, but also the most consistent across buildings, with the lowest standard deviation in R^2 (0.230), whereas the CNN's accuracy varied more widely (standard deviation = 0.312).

4.2. Performance by building type and consumption level

Figure 2 illustrates the mean R^2 by model in five building categories. LGBM outperformed other models in four of the five categories (Education, Lodging, Office, Public). For example, in office buildings, LGBM achieved $R^2 = 0.738$ compared to BiLSTM's $R^2 = 0.631$. Assembly buildings were the exception, where BiLSTM led with $R^2 = 0.576$.

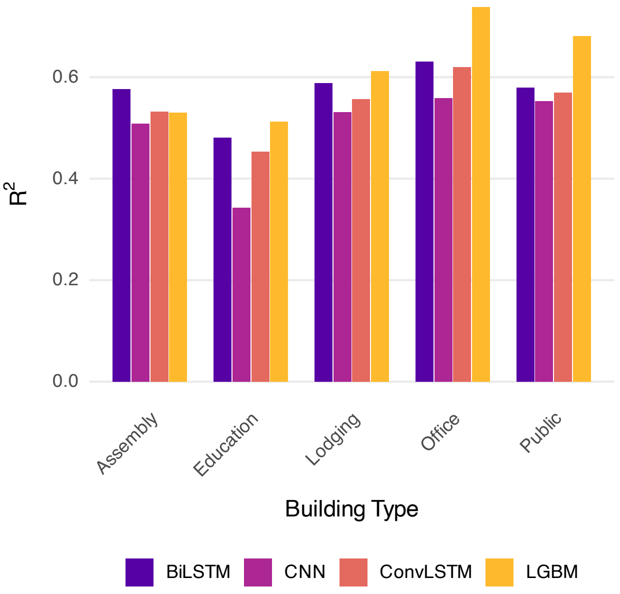

Figure 2: Model R^2 performance by building type

Performance also varied with building energy usage levels (Figure 3). The models achieved their highest R^2 values in medium-to-high consumption buildings with an average consumption between 60 and 120 kWh. For low-usage buildings (<30 kWh), LGBM maintained a relatively strong $R^2 \approx 0.35$, outperforming BiLSTM (0.30) and CNN (0.26). In high-consumption buildings (>120 kWh), LGBM again led with $R^2 \approx 0.66$.

Table 5
Overall performance comparison across all buildings

Model	R^2 Mean	R^2 Std	RMSE Mean	RMSE Std	MAE Mean	MAE Std	Training Time (s)
CNN	0.49	0.31	43.79	104.87	28.89	69.10	8.00
BiLSTM	0.57	0.24	40.01	92.86	25.84	59.67	**5.60**
ConvLSTM	0.55	0.25	40.92	95.71	26.45	61.62	11.06
LGBM	**0.63**	**0.23**	**34.65**	**76.93**	**20.56**	**45.23**	65.49

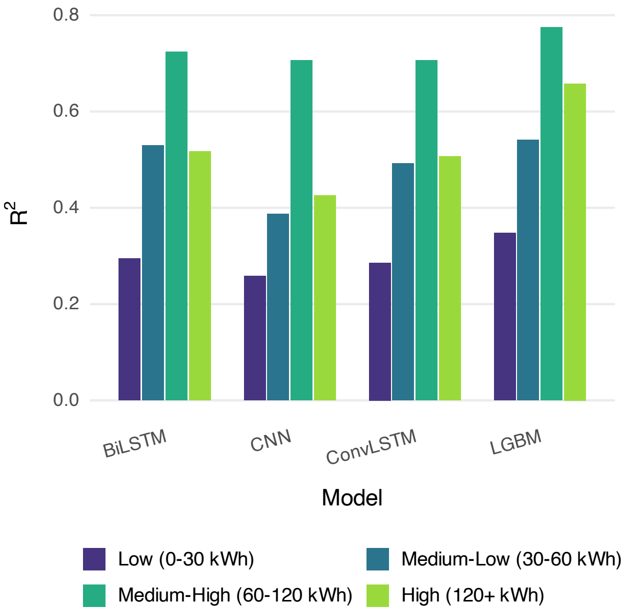

Figure 3: Model R^2 performance by energy usage quartile

4.3. Per-building model performance

Table 6 lists the best model per building based on R^2. LGBM was the top model in 16 of 20 buildings (80%), while BiLSTM was the leader in 4 buildings (20%). The CNN and ConvLSTM models did not outperform other models on any building. LGBM achieved the highest individual R^2 (0.948) on `Hog_office_Richelle` and also performed well in other public and office buildings. BiLSTM excelled on buildings with irregular patterns, such as `Lamb_assembly_Librada` and `Lamb_education_Robin`, indicating its strength in modelling noisy temporal sequences.

4.4. Training time and composite score

Table 5 also reports model training times. BiLSTM had the shortest mean training time (5.6 s), followed by CNN (8.0 s), ConvLSTM (11.1 s) and LGBM (65.5 s). Table 7 presents a composite score that combines R^2 (50%), MAE (30%), and training time (20%) in an effort to balance both predictive performance and efficiency. LGBM led with a score of 0.800, followed by BiLSTM (0.603), ConvLSTM (0.478), and CNN (0.192).

5. Discussion

The results demonstrate that both deep learning models and gradient boosting techniques offer effective solutions for building energy forecasting, with LightGBM showing the most consistent and accurate overall performance. Its ability to capture nonlinear dependencies from structured features such as time-of-day, weather, and occupancy patterns likely contributed to its strong R^2 and low error metrics across most building types and sizes. LGBM was particularly dominant in office, public, and lodging buildings.

BiLSTM performed best in settings with highly irregular or low-magnitude consumption patterns, especially in assembly-type buildings, where occupancy may fluctuate sharply. Its strength in capturing sequential dependencies allowed it to outperform LGBM in those challenging cases. Unlike LGBM, the BiLSTM model rarely suffered catastrophic failures and demonstrated consistent reliability across the dataset.

Despite generally lower performance, the CNN model still achieved meaningful gains over naive baselines, but struggled to compete with models that better leverage temporal information. The ConvLSTM model, while more sophisticated, did not significantly outperform the simpler BiLSTM model in this context, suggesting that convolutional complexity offered limited benefits for univariate time series with engineered features.

From a computational standpoint, BiLSTM was found to be the fastest to train, offering potential advantages for applications requiring frequent model retraining. Although LGBM had higher training times in our implementation, its overall runtime remains acceptable (around one minute per building) and could be reduced with parallelisation or fewer iterations.

Our findings align with previous literature on recurrent neural networks for time series forecasting, yet highlight that properly tuned gradient boosting models can match or exceed deep learning performance in many real-world energy forecasting settings. The composite score results further support LightGBM as a balanced choice when both accuracy and operational efficiency are considered.

The performance of these models has tangible implications for operational deployment in commercial building portfolios. Given its favourable trade-off between accuracy and training time, LightGBM is well suited for *default deployment* scenarios where rapid retraining and scalability are important. For mission-critical applications requiring maximum accuracy, BiLSTM or ConvLSTM architectures

Table 6
Best performing model per building (ranked by R^2)

Building	Type	Mean Energy (kWh)	Best Model	R^2	RMSE	MAE
Hog_office_Richelle	office	85.41	LGBM	0.95	6.01	3.64
Hog_office_Lizzie	office	89.46	LGBM	0.91	9.42	6.95
Peacock_office_Annie	office	76.66	LGBM	0.90	10.55	7.02
Hog_public_Octavia	public	195.51	LGBM	0.86	16.08	10.35
Rat_education_Nellie	education	434.94	LGBM	0.83	66.55	45.32
Robin_office_Shirlene	office	31.44	LGBM	0.79	4.35	3.05
Fox_education_Heriberto	education	37.14	LGBM	0.79	6.73	4.57
Fox_public_Rhonda	public	107.86	LGBM	0.78	18.05	12.39
Peacock_lodging_Sergio	lodging	96.28	LGBM	0.73	10.67	8.05
Robin_lodging_Elmer	lodging	93.68	LGBM	0.68	18.40	10.14
Lamb_assembly_Librada	assembly	31.28	BiLSTM	0.64	23.51	16.59
Hog_lodging_Shanti	lodging	265.71	LGBM	0.64	42.71	24.55
Lamb_education_Robin	education	38.31	BiLSTM	0.59	20.75	15.42
Hog_office_Elsy	office	27.45	LGBM	0.57	5.49	2.32
Rat_assembly_Kimberley	assembly	66.51	BiLSTM	0.51	14.09	7.47
Peacock_lodging_Mathew	lodging	55.15	BiLSTM	0.48	11.27	6.25
Rat_public_Margart	public	43.36	LGBM	0.41	21.90	6.77
Peacock_education_Robbie	education	43.95	LGBM	0.36	9.68	7.59
Hog_office_Joey	office	1198.98	LGBM	0.31	355.61	208.11
Bear_education_Chana	education	21.88	LGBM	0.12	14.91	2.48

Table 7
Composite performance score (R^2: 50%, MAE: 30%, Time: 20%)

Model	R^2	MAE	Time (s)	Score
CNN	0.49	28.89	8.00	0.19
BiLSTM	0.57	25.84	**5.60**	0.60
ConvLSTM	0.55	26.45	11.06	0.48
LGBM	**0.63**	**20.56**	65.49	**0.80**

offer reliable performance, particularly for buildings with irregular usage profiles.

5.1. Limitations

Although the models performed well on most buildings, some facilities exhibited low predictive accuracy ($R^2 < 0.4$) in all methods. These included small educational buildings with high relative volatility and irregular patterns (e.g., Bear_education_Chana, Lamb_education_Robin). Such cases highlight limitations in the available features or the inadequacy of historical trends to capture abrupt behavioural changes. Data preprocessing strategies such as outlier removal, or the incorporation of real-time occupancy or control signals, may improve outcomes in such cases.

Additionally, this study relied on a single train/test split (2016/2017), which limits the assessment of temporal robustness and seasonal adaptation. Cross-validation or rolling-origin experiments could provide a more thorough evaluation of stability over time. The evaluation also focused exclusively on 24-hour-ahead hourly prediction; model rankings may differ for other horizons (e.g., short-term 1-hour ahead or long-term week-ahead forecasts).

Finally, the input features were engineered (e.g., calendar, weather), and models did not directly process raw meter sequences. This approach ensured fair comparison, but may underutilise the representation learning capabilities of deep architectures. Future work should assess models that ingest raw or minimally processed data, including multivariate sequences.

5.2. Future research directions

Several promising avenues remain for future work:

- **Hybrid and ensemble models**: Combining predictions from LGBM and BiLSTM can leverage their complementary strengths, especially in heterogeneous building portfolios.

- **Transfer learning and pretraining**: Recurrent networks trained on large building datasets could be fine-tuned for specific facilities, reducing the need for extensive local training data.

- **Alternative architectures**: Transformers and attention-based models have shown success in time series forecasting and should be evaluated for energy load prediction.

- **Multi-horizon forecasting**: Exploring how model performance varies across different forecast windows (e.g. 1h, 6h, 48h) would reveal each model's temporal strengths.

6. Conclusion

This study conducted a comparative evaluation of four machine learning models – CNN, BiLSTM, ConvLSTM, and LightGBM – for 24-hour ahead energy forecasting across 20 diverse commercial buildings. Using more than 170,000 hourly test samples, we demonstrated that both deep learning and gradient boosting techniques substantially improve prediction accuracy, with all models achieving positive R^2 values and outperforming naive baselines.

Among the models, LightGBM consistently delivered the highest overall performance, achieving the best R^2 in 80% of buildings and yielding the strongest composite score when balancing accuracy, error and training time. Its ability to efficiently capture nonlinear relationships from engineered features makes it a pragmatic and effective choice for deployment across a wide range of building types and consumption profiles.

BiLSTM offered comparable accuracy and emerged as the most reliable model on buildings with irregular or volatile consumption patterns, particularly in assembly and low-demand facilities. Its fast training time and robustness make it a strong alternative when interpretability or temporal dynamics are critical. ConvLSTM and CNN performed reasonably well but did not outperform the simpler BiLSTM or the more interpretable LGBM in most cases.

The findings suggest that model selection should not default to architectural complexity. Instead, it should be informed by building-specific characteristics, operational goals, and infrastructure constraints. For most deployment scenarios, LightGBM offers an optimal balance of performance and efficiency. For high-stakes applications or buildings with complex temporal behaviour, recurrent architectures such as BiLSTM remain highly valuable.

Looking ahead, future research should investigate hybrid ensembles, multi-horizon forecasting, and Transformer-based time series models. Incorporating additional features such as occupancy, controls, or weather forecasts can further enhance the precision of a model. As machine learning continues to integrate into building energy systems, our results provide actionable guidance on selecting forecasting architectures that are accurate, scalable, and deployable.

Reproducibility statement

All data used in this study was obtained from the Building Data Genome 2 dataset [12], available at `https://github.com/buds-lab/building-data-genome-project-2`. The complete codebase used for data preprocessing, model training, and evaluation can be accessed at `https://github.sydney.edu.au/asin2349/THU_0804_SOFT3888`. The experiments were conducted using PyTorch on a Google Colab T4 GPU.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper was prepared as part of the SOFT3888: Software Development Project course at the University of Sydney, under the supervision of Dr. Faham Tahmasebinia. The authors also thank their tutor, Xinrui Zhang, for valuable guidance and support throughout the project.

References

[1] Al-Shargabi, A.A., Almhafdy, A., Ibrahim, D.M., Alghieth, M., Chiclana, F., 2022. Buildings' energy consumption prediction models based on buildings' characteristics: Research trends, taxonomy, and performance measures. Journal of Building Engineering 54, 104577. doi:10.1016/j.jobe.2022.104577.

[2] Chen, G., Lu, S., Zhou, S., Tian, Z., Kim, M.K., Liu, J., Liu, X., 2025. A systematic review of building energy consumption prediction: From perspectives of load classification, data-driven frameworks, and future directions. Applied Sciences 15, 3086. doi:10.3390/app15063086.

[3] Cheng, C.S., Chen, P.W., Jen, H.Y., Wu, Y.T., 2025. A multimodal convolutional neural network framework for intelligent real-time monitoring of etchant levels in PCB etching processes. Mathematics 13, 2804. doi:10.3390/math13172804.

[4] Dong, B., Cao, C., Lee, S.E., 2005. Applying support vector machines to predict building energy consumption in tropical regions. Energy and Buildings 37, 545–553. doi:10.1016/j.enbuild.2004.09.009.

[5] Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM networks, in: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, IEEE. pp. 2047–2052. doi:10.1109/IJCNN.2005.1556215.

[6] Huang, J., Kaewunruen, S., 2023. Forecasting energy consumption of a public building using transformer and support vector regression. Energies 16, 966. doi:10.3390/en16020966.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 3149–3157.

[8] Li, F., Peng, T., Chen, J., Wu, J., Cao, J., Luo, H., Luo, J., Wang, Z., 2025. Prediction and strategies of buildings' energy consumption: A review of modeling approaches and energy-saving technologies. International Journal of Green Energy 22, 2717–2752. doi:10.1080/15435075.2025.2471981.

[9] Li, Q., Meng, Q., Cai, J., Yoshino, H., Mochida, A., 2009. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Conversion and Management 50, 90–96. doi:10.1016/j.enconman.2008.08.033.

[10] Liang, X., Chen, S., Zhu, X., Jin, X., Du, Z., 2023. Domain knowledge decomposition of building energy consumption and a hybrid data-driven model for 24-h ahead predictions. Applied Energy 344, 121244. doi:10.1016/j.apenergy.2023.121244.

[11] Love, F., 2024. NNTikZ – TikZ diagrams for deep learning and neural networks. https://github.com/fraserlove/nntikz. GitHub repository.

[12] Miller, C., Kathirgamanathan, A., Picchetti, B., Arjunan, P., Park, J.Y., Nagy, Z., Raftery, P., Hobson, B.W., Shi, Z., Meggers, F., 2020. The building data genome project 2: Energy meter data from the ASHRAE great energy predictor III competition. Scientific Data 7, 368. doi:10.6084/m9.figshare.13033847.

[13] Moveh, S., Merchán-Cruz, E.A., Abuhussain, M., Dodo, Y.A., Alhumaid, S., Alhamami, A.H., 2025. Deep learning framework using transformer networks for multi-building energy consumption prediction in smart cities. Energies 18, 1468. doi:10.3390/en18061468.

[14] Neale, A., Kummert, M., Bernier, M., 2022. Development of a bottom-up white-box building stock energy model for single-family dwellings. Journal of Building Performance Simulation 15, 735–756. doi:10.1080/19401493.2022.2082531.

[15] Olu-Ajayi, R., Alaka, H., Owolabi, H., Akanbi, L., Ganiyu, S., 2023. Data-driven tools for building energy consumption prediction: A review. Energies 16, 2574. doi:10.3390/en16062574.

[16] Pachano, J.E., Nuevo-Gallardo, C., Fernández Bandera, C., 2025. An empirical comparison of a calibrated white-box versus multiple LSTM black-box building energy models. Energy and Buildings 333, 115485. doi:10.1016/j.enbuild.2025.115485.

[17] Pecan Street Inc., 2025. Pecan street dataport. https://www.pecanstreet.org/dataport/.

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[19] Russell, S., 1991. Inductive learning by machines. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 64, 37–64. URL: https://www.jstor.org/stable/4320245.

[20] Schmitt, M., Ahmadi, S.A., Xu, Y., Taşkin, G., Verma, U., Sica, F., Hänsch, R., 2023. There are no data like more data: Datasets for deep learning in earth observation. IEEE Geoscience and Remote Sensing Magazine 11, 63–97. doi:10.1109/MGRS.2023.3293459.

[21] Sunder, R., R., S., Paul, V., Punia, S.K., Konduri, B., Nabilal, K.V., Lilhore, U.K., Lohani, T.K., Ghith, E., Tlija, M., 2024. An advanced hybrid deep learning model for accurate energy load prediction in smart buildings. Energy Exploration & Exploitation 42, 2241–2269. doi:10.1177/01445987241267822.

[22] Tsanas, A., Xifara, A., 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings 49, 560–567. doi:10.1016/j.enbuild.2012.03.003.

[23] Wong, S., Wan, K.K., Lam, T.N., 2010. Artificial neural networks for energy analysis of office buildings with daylighting. Applied Energy 87, 551–557. doi:10.1016/j.apenergy.2009.06.028.

[24] Yang, J., Rivard, H., Zmeureanu, R., 2005. Building energy prediction with adaptive artificial neural networks, in: Proceedings of Building Simulation 2005: 9th Conference of IBPSA, IBPSA, Montréal, Canada. pp. 1401–1408. doi:10.26868/25222708.2005.1401-1408.

[25] Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., Livingood, W., 2021. A review of machine learning in building load prediction. Applied Energy 285, 116452. doi:10.1016/j.apenergy.2021.116452.

[26] Zhang, Y., Wang, D., Wang, G., Xu, P., Zhu, Y., 2025. Data-driven building load prediction and large language models: Comprehensive overview. Energy and Buildings 326, 115001. doi:10.1016/j.enbuild.2024.115001.

[27] Zhao, H.x., Magoulès, F., 2012. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews 16, 3586–3592. doi:10.1016/j.rser.2012.02.049.