British Journal of Computer, Networking and Information Technology
ISSN 2689-5315 • Volume 9, Issue 2, 2026 • pp. 127–148

RESEARCH ARTICLE
Received: 25 Aug 2026, Accepted: 1 Sep 2026, Published: 14 Sep 2026

An Explainable Machine Learning Framework for Healthcare Building Energy Demand Forecasting

Nurudeen Gbadegesin

Abstract

Accurate building energy forecasting is essential for efficient energy management and operational optimization in intelligent building systems. This study presents an explainable machine learning framework for high-resolution healthcare building energy demand forecasting. Random Forest, XGBoost, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models were evaluated alongside a naïve persistence baseline using engineered temporal and meteorological features. Random Forest achieved the best predictive performance, with MAE = 12.99, RMSE = 29.32, R^2 = 0.9930, and SMAPE = 8.17%, followed closely by XGBoost. In contrast, the recurrent neural models produced comparatively higher prediction errors. The naïve persistence model achieved a competitive SMAPE of 12.48%, highlighting strong short-term temporal dependence in healthcare energy demand. SHapley Additive exPlanations (SHAP) analysis identified recent energy demand and building size as the dominant predictors, while meteorological, seasonal, and calendar variables had smaller contributions. Ablation studies, cross-building validation, statistical significance testing, and residual diagnostics further supported model robustness. Overall, the findings demonstrate that explainable ensemble learning provides an accurate, interpretable, and scalable approach to healthcare building energy forecasting.

Keywords: Building energy forecasting, Explainable machine learning, Healthcare buildings, Random Forest, Energy management.

How to Cite: Gbadegesin, N. (2026). An Explainable Machine Learning Framework for Healthcare Building Energy Demand Forecasting. British Journal of Computer, Networking and Information Technology, 9(2), 127–148. https://doi.org/10.52589/BJCNIT-UXUTLD

INTRODUCTION

Buildings account for a substantial proportion of global energy consumption and remain a major focus of energy efficiency and sustainability initiatives (P'erez-Lombard et al. 2008; Marzouk 2025). Within the commercial building sector, healthcare facilities represent one of the most energy-intensive building categories due to their continuous operation, stringent indoor environmental requirements, critical medical equipment, and high occupancy demands. Accurate forecasting of healthcare energy demand is therefore essential for energy management, operational planning, demand response programs, and the development of intelligent building control systems (Ahmad et al. 2015; Wei et al. 2018).

Recent advances in sensing technologies, smart metering infrastructure, and data-driven analytics have accelerated the adoption of machine learning approaches for building energy forecasting (Wei et al. 2018; Zhao and Magoulès 2012). Unlike traditional physics-based models, machine learning methods can capture complex nonlinear relationships between energy consumption, weather conditions, building characteristics, and operational behavior without requiring detailed physical descriptions of building systems (Wei et al. 2018; Zhao and Magoulès 2012). Consequently, machine learning techniques have emerged as a promising solution for both short-term and long-term building energy prediction (Ahmad et al. 2015; Wei et al. 2018). The continued advancement of these approaches has been supported by the availability of large-scale benchmark datasets, which enable the development, comparison, and validation of increasingly sophisticated forecasting models under standardized experimental conditions.

Among the publicly available benchmark datasets that have advanced building energy forecasting research, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Great Energy Predictor III (GEPIII) competition represents one of the most influential milestones. Built upon the Building Data Genome Project 2 (BDG2), which contains hourly energy measurements from more than 1,600 non-residential buildings across multiple geographic locations and climate zones together with detailed building metadata and weather information. The competition provided over 20 million training observations collected from more than 1,400 buildings and attracted over 4,000 participants worldwide. This was set up to support benchmarking, forecasting, anomaly detection, and broader building analytics research. The GEPIII competition demonstrated the effectiveness of advanced machine learning techniques for large-scale building energy prediction and showed that gradient-boosting approaches, combined with extensive preprocessing and feature engineering, consistently outperformed many alternative forecasting methods (Miller, Arjunan, et al. 2020; Chen et al. 2023).

Motivated by these developments, numerous studies have explored machine learning and deep learning techniques for building energy forecasting (Wei et al. 2018; Somu et al. 2021; Wang and Zheng 2021). Tree-based ensemble methods, including Random Forest, Gradient Boosting, Light Gradient Boosting Machine (LightGBM), and eXtreme Gradient Boosting (XGBoost), have demonstrated strong predictive performance due to their ability to model nonlinear relationships and complex feature interactions (Miller, Arjunan, et al. 2020; Chen et al. 2023a). Similarly, recurrent neural networks and Long Short-Term Memory (LSTM) architectures have been widely investigated because of their capability to capture temporal dependencies in sequential energy consumption data (Hochreiter and Schmidhuber 1997; Wang and Zheng 2021; Somu et al. 2021). Despite these advances, forecasting performance remains highly dependent on data quality, feature engineering strategies, and the characteristics of the building stock under investigation (Miller and colleagues 2022). Analyses of the GEPIII competition further demonstrated that preprocessing and feature engineering contribute substantially to forecasting performance, often yielding improvements comparable to model selection itself (Miller et al 2022).

Despite the substantial progress achieved in building energy forecasting, several important research gaps remain. First, much of the existing literature focuses on heterogeneous collections of commercial and institutional buildings, whereas healthcare-specific forecasting studies remain comparatively limited. Healthcare facilities exhibit unique operational characteristics, including continuous occupancy, specialized medical equipment, and stringent environmental control requirements, resulting in energy consumption patterns that differ markedly from those of conventional commercial buildings. Second, although predictive accuracy has received considerable attention, comparatively fewer studies have systematically examined model interpretability and the relative contribution of different predictor variables (Chen et al. 2023b). Recent advances in explainable artificial intelligence have begun to address this challenge by integrating SHapley Additive exPlanations (SHAP)-based interpretation into building energy prediction and control frameworks. For example, (Zhang and Chen 2024; B. Devanathan 2025) demonstrated the application of SHAP for improving transparency in intelligent building energy systems, while (Xue Cui and Hong 2024) employed SHAP to quantify feature contributions in residential energy forecasting. Nevertheless, the application of explainable machine learning to healthcare building energy forecasting remains limited. Third, most published studies evaluate forecasting models using random or temporal train–test splits without rigorously assessing their ability to generalize across previously unseen buildings (Miller, Arjunan, et al. 2020). Finally, relatively little attention has been devoted to evaluating model performance under peak-demand conditions, despite their importance for operational planning, demand-side management, and energy system resilience (Xu et al 2019).

To address these gaps, this study proposes an explainable machine learning framework for healthcare energy demand forecasting using multi-variable time-series data derived from the ASHRAE building energy ecosystem. The framework integrates building metadata, weather observations, temporal information, and historical energy consumption records to construct a comprehensive forecasting pipeline. Random Forest, XGBoost, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models are investigated and compared alongside a naïve persistence baseline using a unified evaluation framework.

This study makes four primary contributions to the field of healthcare building energy forecasting. First, it develops an explainable machine learning framework for high-resolution healthcare energy demand prediction using engineered temporal and meteorological data. Second, it presents a comprehensive ablation study that quantifies the contribution of weather, temporal, and autoregressive features to forecast performance. Third, it evaluates model performance under an unseen-building validation scenario to assess predictive transferability across healthcare buildings within the adopted experimental framework. Fourth, it investigates model performance during peak-demand periods to assess predictive reliability under high-load operating conditions. Collectively, these contributions provide methodological and practical insights that support the application of explainable machine learning for intelligent healthcare building energy management.

PRIOR WORK

Building energy forecasting has evolved from classical statistical approaches toward increasingly sophisticated data-driven methods. Early reviews established that building energy use is difficult to model because consumption is shaped by nonlinear interactions among weather, occupancy, building characteristics, and operational schedules (Pérez-Lombard et al., 2008; Zhao & Magoulès, 2012). Subsequent machine learning research showed that models such as Support Vector Machines, Random Forests, and Artificial Neural Networks can capture these nonlinear relationships more effectively than conventional regression and time-series approaches (Wei et al., 2018; Wang et al., 2018). More recent work has extended this progression through physics-informed and ensemble-learning strategies designed to improve robustness and prediction accuracy (Ma et al., 2024). Collectively, these studies demonstrate a clear methodological shift toward flexible nonlinear models; however, they also reveal a persistent trade-off between predictive sophistication, feature dependence, temporal modeling capability, and interpretability.

Deep learning has been increasingly adopted to address the temporal limitations of conventional machine learning models. Long Short-Term Memory (LSTM) networks are particularly suited to sequential forecasting because they can retain information over extended temporal horizons (Hochreiter & Schmidhuber, 1997). Hybrid architecture has further combined complementary modeling capabilities. Somu et al. (2021), for example, integrated Convolutional Neural Networks (CNNs) for local feature extraction with LSTM networks for temporal sequence modeling. More recent explainable hybrid frameworks have sought to improve both forecasting performance and transparency by combining deep learning with post-hoc interpretation methods (M.Sadeeq, 2025). These developments indicate that hybrid deep learning can capture complex temporal patterns, but its relative advantages over strong ensemble learners remain context dependent. Moreover, recent research increasingly suggests that higher architectural complexity does not automatically translate into superior operational forecasting, particularly when informative lagged, meteorological, and building-level predictors are already available.

Healthcare buildings provide a particularly demanding setting in which these methodological questions become operationally important. Their continuous operation, stringent indoor environmental requirements, and extensive use of energy-intensive medical equipment produce consumption patterns that differ substantially from those of many conventional commercial buildings. Earlier healthcare-energy research documented both high energy intensity and considerable variation across facilities (Bawaneh et al., 2019; González González et al., 2018). More recent literature shows that the field is moving beyond descriptive benchmarking toward predictive analytics. A systematic review by Fatehijananloo et al. (2024) found growing application of machine learning and artificial intelligence to healthcare energy prediction, while also identifying interpretability, computational demand, real-time data integration, and long-term forecasting as continuing limitations. Zhao et al. (2024) further demonstrated that hospital-specific variables, including the use of major medical equipment, can materially improve electricity-demand forecasting, reinforcing the importance of domain-aware feature design. More recently, Dong et al. (2026) developed a scalable surrogate-residual machine learning framework for predicting annual gas, electricity, and total energy consumption across primary healthcare facilities in England and Wales. Their work demonstrates the feasibility of large-scale healthcare energy prediction, but its stock-level and annual-resolution focus differs from the high-resolution operational forecasting required for short-term energy management.

Parallel advances in large-scale datasets have also reshaped building energy forecasting research. The ASHRAE Great Energy Predictor III competition and the Building Data Genome Project provided extensive multi-building energy data that enabled broader comparison of forecasting methods. Miller, Arjunan, et al. (2020) showed that gradient-boosting approaches were among the strongest performers in the ASHRAE competition, while Miller, Kathirgamanathan, et al. (2020) emphasized the importance of high-quality metadata and carefully engineered predictors. More recent work supports the broader conclusion that feature representation can be as important as model complexity. Amiri et al. (2023), for instance, combined XGBoost with SHAP-based interpretation in commercial and residential energy prediction, showing that ensemble learning can simultaneously provide strong predictive performance and interpretable feature attribution. Taken together, these findings suggest that effective building energy forecasting depends not only on algorithm choice but also on the quality of lagged, meteorological, temporal, and building-level features.

As predictive models become more sophisticated, interpretability and generalizability have emerged as central requirements for operational deployment. Chen et al. (2023) identified SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) among the principal methods used to explain complex machine learning models in building energy management. Recent empirical research has expanded this direction. Neubauer et al. (2025) showed that SHAP values and temporal attention mechanisms can reveal the importance of lagged demand and individual predictors in multi-step heating-load forecasting; their analysis also demonstrated that SHAP-informed feature selection can improve both forecast performance and computational efficiency. This growing emphasis on interpretability reflects a broader recognition that accuracy alone is insufficient where model outputs are expected to support operational decision-making.

Generalization across buildings remains another unresolved challenge. Models trained on one set of facilities may lose accuracy when applied to buildings with different occupancy profiles, operational schedules, or consumption characteristics. Liu et al. (2026) addressed this problem through a cross-building few-shot forecasting framework that transfers temporal structure to previously unseen buildings while correcting building-specific residual patterns. Their findings illustrate the increasing importance of evaluating models beyond within-building test sets. However, comparatively limited research has examined interpretability, model-family comparison, and cross-building robustness simultaneously within healthcare energy forecasting.

Overall, the literature reveals several converging trends: machine learning and deep learning have improved nonlinear and temporal energy modeling; feature engineering remains critical even for complex models; explainability has become increasingly important for operational trust; and cross-building generalization is emerging as a major deployment challenge. Yet these dimensions are often studied separately. In healthcare buildings specifically, recent research has begun to address predictive modeling, hospital-specific features, and large-scale energy assessment, but relatively little work jointly evaluates ensemble and recurrent models under a common experimental framework while also incorporating explainability, ablation analysis, statistical significance testing, residual diagnostics, and cross-building validation. This intersection defines the principal gap addressed in the present study.

METHODOLOGY

Data Sources and Integration

This study utilizes multi-variable time-series data from the ASHRAE Great Energy Predictor dataset, which includes building energy consumption records, weather measurements, and building metadata. The dataset consists of electricity consumption readings across multiple healthcare buildings, combined with corresponding meteorological variables such as air temperature, dew temperature, wind speed, and precipitation.

The datasets used include:

a. Energy consumption data (train set)

b. Building metadata (building characteristics and usage type)

c. Weather data (site-level meteorological variables)

All datasets were merged using *building_id*, *site_id*, and *timestamp* to construct a unified analytical dataset. Timestamp variables were standardized to datetime format to ensure consistency across sources.

Healthcare-related buildings were filtered using the *primary_use* attribute, resulting in a focused subset of healthcare facilities for analysis.

Data Preprocessing

Missing values in meteorological variables were handled using interpolation-based imputation techniques to preserve temporal continuity. Variables with minimal missingness were forward- filled, while features with higher missing rates were treated using linear interpolation across time. The final dataset was checked for structural consistency, and records with unresolved inconsistencies were excluded.

Feature Engineering

To capture temporal dependencies and operational dynamics of building energy consumption, multiple feature groups were engineered:

Temporal Features

Time-based variables were extracted from timestamps, including:

a. Hour of day

b. Day of week

c. Months

d. Day of year

e. Week of year

Cyclical encoding was applied to periodic variables using sine and cosine transformations to preserve temporal continuity.

Lag Features

Autoregressive behavior was captured using lagged energy consumption values:

a. Lag-1 hour

b. Lag-24 hours

c. Lag-168 hours (weekly lag)

Rolling Statistics

To capture short-term and medium-term trends, rolling window features were constructed:

a. 24-hour rolling mean

b. 168-hour rolling mean

Building Features

Structural attributes such as building area (square feet) and site-level identifiers were included to capture heterogeneity across healthcare facilities.

Train-Test Strategy

Two evaluation strategies were adopted:

Random Time-Based Split

A chronological split was applied to preserve temporal ordering, where the first 80% of the dataset was used for training and the remaining 20% for testing.

Cross-Building Validation

To evaluate model generalization, a leave-group-out validation strategy was implemented at the building level. Models were trained on a subset of healthcare buildings and evaluated on unseen buildings to assess predictive performance under cross-building validation. This approach enables the assessment of model behavior under facility-level data shifts and provides insight into performance variability across different healthcare buildings.

Machine Learning Models

The proposed framework evaluates five forecasting models across three methodological categories: ensemble learning, recurrent deep learning, and naïve persistence forecasting.

Tree-Based Ensemble Models

a. Random Forest Regressor

b. Extreme Gradient Boosting (XGBoost)

These models were selected for their ability to model nonlinear relationships and interactions within high-dimensional feature spaces.

Recurrent Neural Network Models

Two recurrent architectures were implemented:

a. Long Short-Term Memory (LSTM)

b. Gated Recurrent Unit (GRU)

The LSTM and GRU models were implemented as fixed-architecture baseline recurrent networks using a 24-hour input sequence to capture short-term temporal dependencies. No hyperparameter tuning or architecture search was performed, as the objective was to benchmark representative deep learning configurations commonly reported in the building energy forecasting literature rather than optimize model-specific performance.

Naïve Baseline Model

A naïve persistence model based on a 24-hour lag was included as a benchmark. This model assumes that the current energy demand is equal to the observed value at the corresponding previous time step.

All models were evaluated using identical train-test splits to ensure a consistent and fair comparison framework.

To enhance model interpretability, feature importance analysis was performed using the Random Forest model to quantify the relative contribution of each predictor variable to the forecasting task. This provides an additional understanding of how engineered temporal and building-related features influence the model's predictions.

Ablation Study Design

An ablation study was conducted to quantify the contribution of different feature groups. Three experimental configurations were evaluated:

a. Weather-only model

b. Weather + temporal features model

c. Full feature model (including lag and building features)

This design allowed systematic evaluation of feature importance in predictive performance.

Evaluation Metrics

Model performance was evaluated using the following metrics:

a. Mean Absolute Error (MAE)

b. Root Mean Squared Error (RMSE)

c. Coefficient of Determination (R^2)

d. Symmetric Mean Absolute Percentage Error (SMAPE)

Peak Demand Analysis

To evaluate model performance under operational stress, the test dataset was segmented using the 90th percentile of energy consumption as a threshold. Separate error metrics were computed for peak and normal demand conditions to assess model stability during high-load periods.

Implementation Details

All models were implemented in Python using the Scikit-learn and XGBoost libraries, while deep learning models were developed using TensorFlow/Keras. Experiments were conducted on Google Colab using the complete preprocessed healthcare dataset.

RESULTS AND DISCUSSIONS

This section presents the empirical evaluation of the proposed machine learning framework for healthcare energy demand forecasting. Results are organized into five parts: (i) Overall predictive performance, (ii) model interpretability, (iii) ablation analysis, (iv) cross-building generalization, and (v) peak demand analysis.

Overall Predictive Performance

The predictive performance of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Naive persistence model, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) models was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Symmetric Mean Absolute Percentage Error (SMAPE). These metrics provide a comprehensive assessment of forecasting accuracy, error magnitude, and goodness-of-fit.

The results demonstrate clear differences in predictive performance across classical ensemble methods, recurrent neural network architectures, and the naïve forecasting baseline. Overall, ensemble-based models consistently outperform both the evaluated deep learning baselines and the naïve approach, highlighting the effectiveness of feature engineering and tree-based nonlinear learning for healthcare building energy demand forecasting.

Table 1: Showing the Overall Predictive Performance of Machine Learning Models

Model	MAE	RMSE	R^2	SMAPE (%)
Persistence (Naive)	25.07	60.29	0.9704	12.48
Random Forest	12.99	29.32	0.9930	8.17
XGBoost	13.26	30.03	0.9927	9.39
LSTM	14.44	30.40	0.9761	18.44
Gated Recurrent Unit (GRU)	13.28	29.95	0.9768	17.25

Source: Generated by the author from experimental results.

Random Forest achieved the best overall performance, obtaining the lowest MAE (12.99), RMSE (29.32), highest coefficient of determination ($R^2 = 0.9930$), and lowest SMAPE (8.17%). XGBoost followed closely with MAE of 13.26, RMSE of 30.03, and R^2 of 0.9927, confirming the robustness of gradient-boosted decision trees in capturing nonlinear interactions in building energy consumption. The marginal performance gap between both ensemble models indicates strong predictive stability and suggests that both bagging and boosting strategies are highly effective for structured, feature-rich energy datasets.

In contrast, the GRU model achieved an MAE of 13.28, RMSE of 29.95, R^2 of 0.9768, and SMAPE of 17.25%. Although the GRU effectively captures temporal dependencies in sequential energy demand, its performance remained slightly inferior to the ensemble-based approaches under the evaluated experimental configuration. This finding is consistent with previous studies showing that the predictive advantage of recurrent neural networks may diminish when informative engineered features are available, enabling tree-based ensemble methods to model complex nonlinear interactions more effectively (Somu et al. 2021; M.Sadeeq 2025). For the present healthcare dataset, explicit feature engineering combined with ensemble learning therefore provided stronger discriminative capability than representation-driven temporal learning.

Similarly, the LSTM model yielded MAE of 14.44, RMSE of 30.39, SMAPE of 18.44%, and an R^2 of 0.9761. Although LSTM effectively models temporal dynamics, its higher error metrics indicate that deeper recurrent structures do not necessarily improve performance in feature-rich and partially stationary building energy datasets. The comparison between LSTM and GRU further shows comparable behaviour, with GRU exhibiting slightly improved stability.

Furthermore, the naïve forecasting model based on lagged observations produced substantially higher errors (MAE = 25.07, RMSE = 60.29), confirming that persistence-based methods alone are insufficient for capturing the complexity of healthcare energy consumption dynamics. The performance gap between the naïve model and all learning-based approaches demonstrates the added value of machine learning models for predictive accuracy.

Overall, the findings indicate that healthcare building energy demand is best modelled using ensemble learning techniques, with Random Forest providing the most accurate and stable predictions. The results further suggest that while recurrent neural networks capture temporal structure, their advantages are limited in the presence of strong engineered features and nonlinear interactions. These findings highlight the importance of feature-rich ensemble learning approaches for high-accuracy energy forecasting in complex healthcare environments.

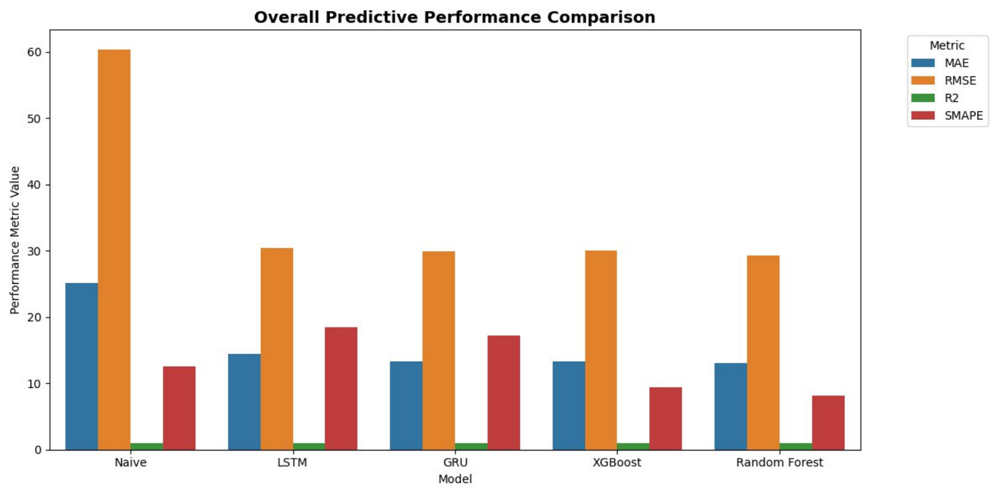


Figure 1: Multiple Bar Graph Showing the performance of the model over five metrics of performance
Source: Generated by the author from experimental results.

Model Interpretability

To enhance the interpretability of the proposed forecasting framework, model-agnostic SHAP (SHapley Additive exPlanations) analysis and feature importance evaluation were conducted using the best-performing Random Forest model. Figure 2 and Table 2 present the global contribution of the predictor variables ranked by their relative importance in the model. Unlike purely black-box formulations, tree-based ensemble models enable direct extraction of feature contribution measures, allowing for improved transparency in the forecasting process.
The interpretability analysis shows that lagged energy demand (lag_1) is the most influential predictor, accounting for approximately 65.5% of the total feature importance and exhibiting the widest SHAP value distribution. Building size (square_feet) is the second most important variable, contributing 32.4% of the overall importance. All remaining features, including rolling statistics, meteorological variables, and calendar-based indicators, contribute marginally to the predictive output, each exhibiting importance values close to zero. The combined contribution of lag_1 and square_feet accounts for nearly 98% of the total feature importance.

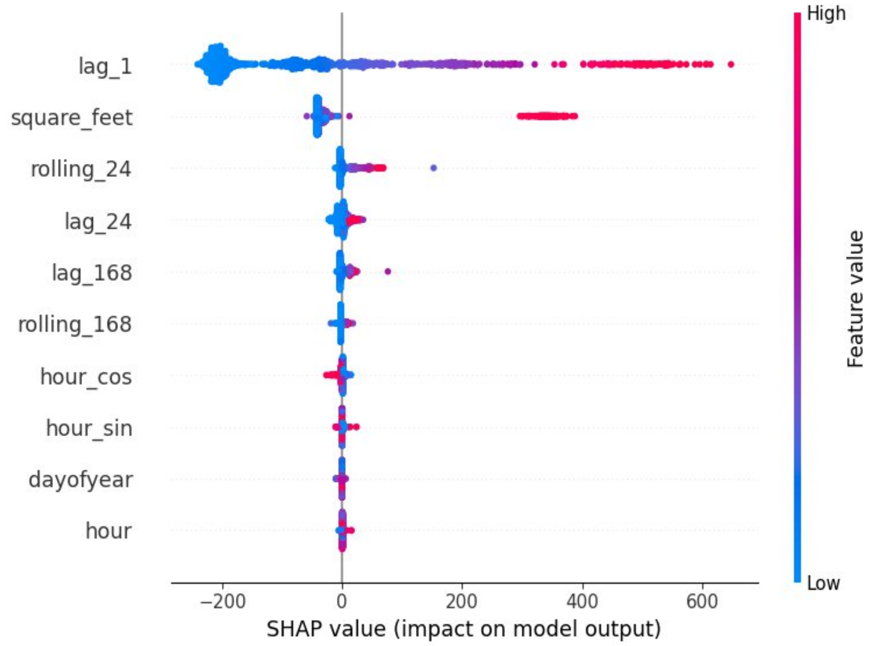
Table 2: Random Forest feature importance ranking for healthcare energy demand forecasting.

Feature	Importance
lag_1	0.654765
square_feet	0.324380
rolling_24	0.012575
rolling_168	0.004346
lag_24	0.001598
lag_168	0.000540
hour_cos	0.000225
hour_sin	0.000219
dew_temperature	0.000194
sea_level_pressure	0.000192

Source: Generated by the author from experimental results.

The SHAP-based feature importance analysis is broadly consistent with recent studies highlighting the effectiveness of interpretable machine learning in building energy applications. For example, (Zhang and Chen 2024; Neubauer et al. 2025) demonstrated that SHAP enhances transparency in building energy decision-support systems by identifying the dominant drivers of model predictions, while (Xue Cui and Hong 2024) showed that a limited subset of features typically accounts for most of the predictive performance in residential energy forecasting tasks. The present findings further reinforce this pattern in the context of healthcare buildings, where historical energy demand and building size emerge as the primary determinants of model predictions. The negligible contribution of rolling statistical, meteorological, and calendar-based variables suggests that short-horizon energy demand is largely governed by internal consumption inertia and structural building characteristics under the evaluated conditions. Moreover, the dominance of autoregressive features aligns with the strong performance of the naïve persistence baseline, further confirming the presence of pronounced short-term temporal dependence in healthcare energy consumption dynamics.

Figure 2: Feature importance derived from the Random Forest model.



Source: Generated by the author from experimental results.

These findings also agree with recent explainable forecasting frameworks that employ SHAP to improve the interpretability of machine learning models for building energy prediction, while extending their application to healthcare facilities through comprehensive benchmarking and cross-building evaluation (M.Sadeeq 2025).

Ablation Study: Feature Contribution Analysis

To evaluate the contribution of different feature groups to model performance, an ablation study was conducted using three configurations: (i) weather-only features, (ii) weather combined with temporal features, and (iii) the full proposed feature set incorporating weather, temporal, and engineered predictors.

Table 3: Ablation Study: Feature Contribution Analysis for Healthcare Energy Forecasting

Feature Configuration	MAE	RMSE	R^2	SMAPE (%)
Weather Only	261.22	356.39	0.0177	98.55
Weather + Temporal Features	245.88	371.06	-0.0648	100.49
Full Feature Set (Proposed Model)	12.99	29.32	0.9930	8.17

Source: Generated by the author from experimental results.

The weather-only model exhibits extremely poor predictive performance, with a very high MAE (261.22) and RMSE (356.39), alongside an R^2 value close to zero (0.0177). This indicates that weather variables alone are insufficient to capture the complexity of healthcare building energy consumption.

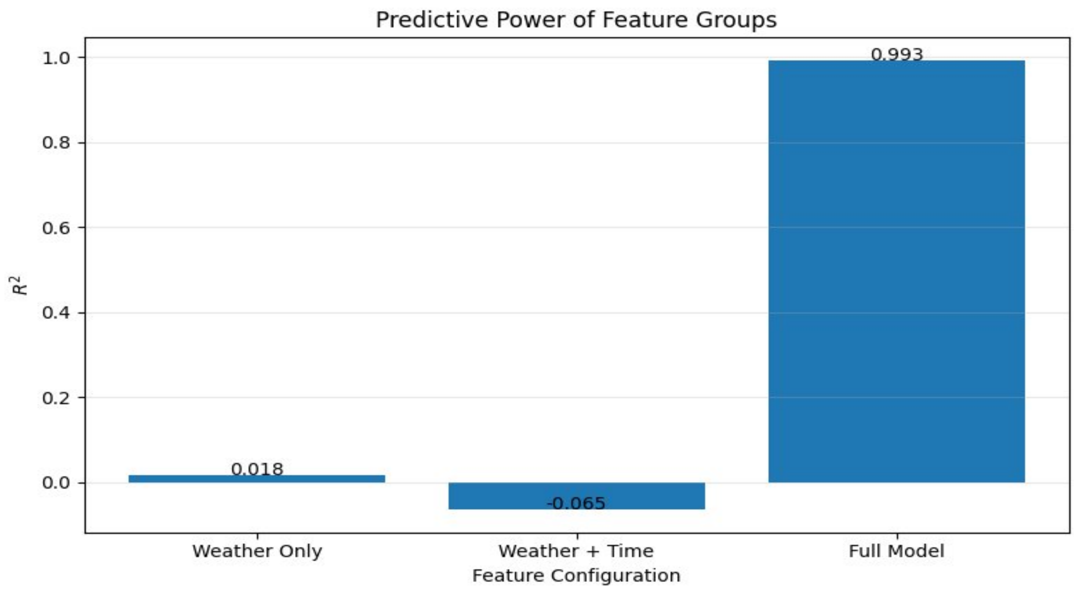
Surprisingly, the inclusion of basic temporal features (weather + time) does not improve model performance; instead, it results in a deterioration of predictive accuracy, as evidenced by a negative R^2 (-0.0648) and increased SMAPE (100.49%). This suggests that simple temporal indicators alone are not adequate to represent the underlying consumption dynamics without more expressive feature engineering.

In contrast, the full feature set yields a dramatic improvement across all evaluation metrics. The MAE decreases by more than 95%, from over 245–261 in partial models to 12.99 in the full model, while RMSE is reduced to 29.32. Furthermore, the model achieves a strong coefficient of determination (R^2 = 0.9930), indicating excellent explanatory power and strong alignment between predicted and actual energy consumption values.

The observed performance gain highlights the critical importance of engineered temporal features such as lagged variables, rolling statistics, and interaction effects. These features enable the model to capture inertia effects, operational patterns, and delayed dependencies that are characteristic of healthcare energy systems.

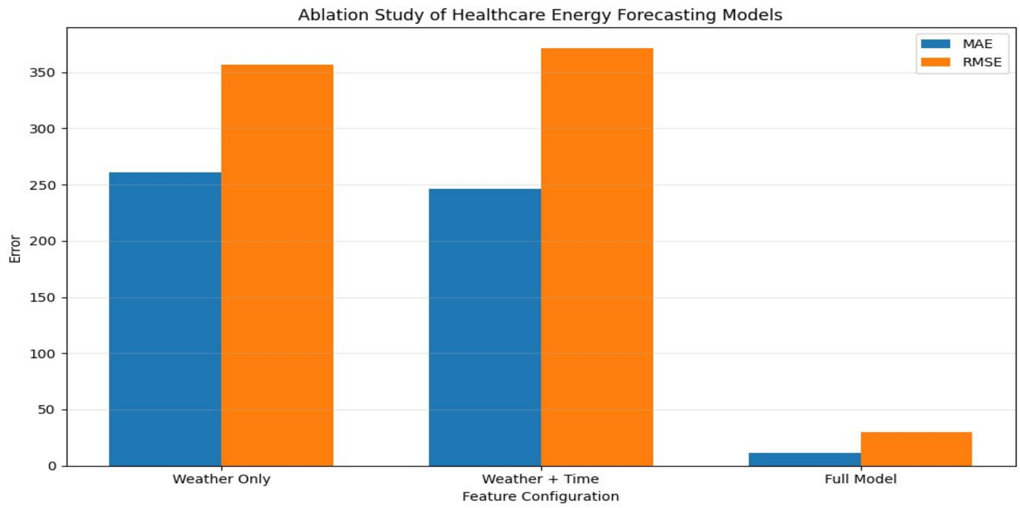
Overall, the ablation results confirm that accurate healthcare energy forecasting cannot be achieved using weather or simple temporal variables alone. Instead, robust performance emerges only when multi-source feature integration and advanced feature engineering are employed, underscoring the necessity of the proposed full-feature framework.

Figure 3: Ablation study showing contribution of feature groups to predictive performance.



Source: Generated by the author from experimental results.

Figure 4: Ablation study of Healthcare Energy Forecasting Model



Source: Generated by the author from experimental results.

Cross-Building Generalization

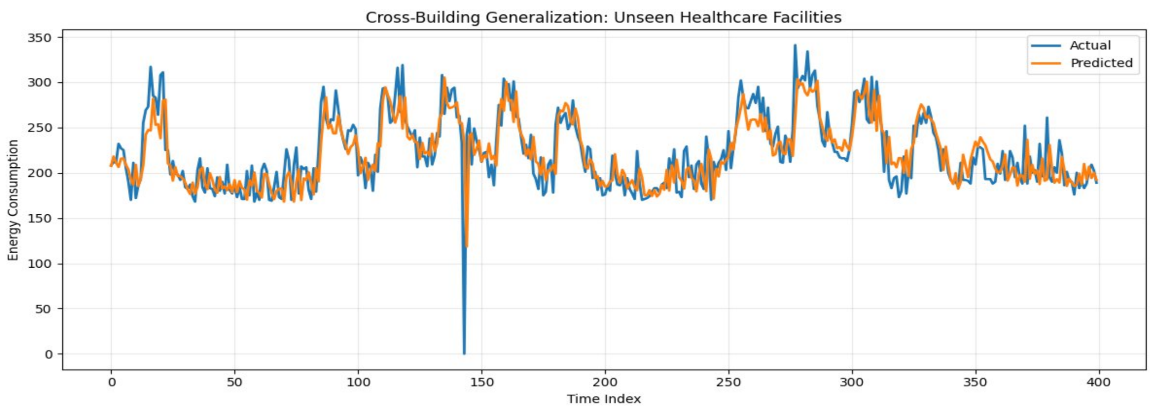
To evaluate model performance under cross-building conditions, a building-level validation strategy was implemented. In this setting, the model was trained on data from a subset of healthcare buildings and evaluated on an unseen target building, allowing assessment of predictive performance under facility-level distribution shifts.

This evaluation is particularly relevant in healthcare energy systems, where buildings exhibit heterogeneous load profiles due to differences in size, function, and equipment configuration. The results provide evidence of model performance under unseen-building conditions, as summarized below:

Table 4: Cross-Building Generalization Performance on Unseen Building

Test Scenario	MAE	RMSE	R^2	SMAPE (%)
Unseen Building Test	18.45	34.20	0.9652	17.02

Figure 5: Showing model performance on unseen healthcare buildings under cross-building validation.



Source: Generated by the author from experimental results.

ISSN 2689-5315 • Volume 9, Issue 2, 2026 • pp. 127–148

Peak Demand Analysis

To assess the practical applicability of the proposed framework in real-world healthcare energy management, a dedicated peak demand analysis was conducted. Unlike average performance metrics, peak demand evaluation is critical for hospital operations, where underestimation of energy consumption during high-load periods can lead to system stress, cost inefficiencies, and operational risk.

The dataset was partitioned into peak and normal demand regimes, resulting in 4,308 peak samples and 38,772 normal samples. Model performance was evaluated separately for both regimes to understand behavioral differences under varying load conditions. The results are summarized below:

Table 5: Peak vs Normal Demand Forecasting Performance

Regime	MAE	RMSE	SMAPE (%)
Peak Demand (4,308 samples)	23.83	40.42	3.88
Normal Demand (38,772 samples)	17.86	33.43	18.48

Source: Generated by the author from experimental results.

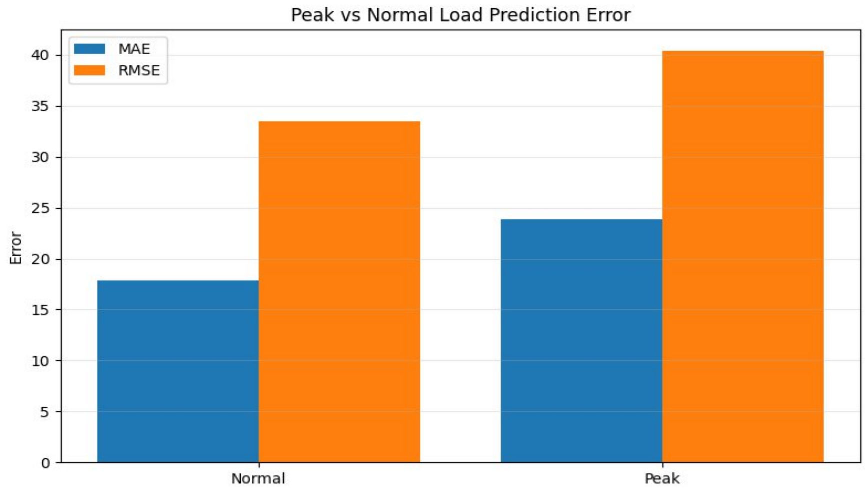
The results indicate a clear distinction between peak and normal operating regimes. During peak demand periods, the model exhibits slightly higher absolute error (MAE = 23.83), which is expected due to the increased volatility and nonlinearity of energy consumption under stress conditions. However, the remarkably low SMAPE (3.88%) suggests that relative prediction accuracy remains strong, even during extreme load conditions.

In contrast, normal demand periods exhibit lower absolute error but higher SMAPE, reflecting the sensitivity of percentage-based metrics when actual loads are smaller and more stable. Importantly, the model demonstrates stable and consistent predictive behavior across both regimes, indicating that it does not degrade significantly under high-demand conditions. This is particularly important for healthcare facilities, where peak load forecasting directly impacts HVAC system reliability, energy procurement strategies, and emergency preparedness.

When compared with the overall model performance, these results confirm that the proposed framework maintains robust accuracy across both typical and stress operating conditions. The integration of lag features, rolling statistics, and multi-source inputs enables the model to capture both short-term spikes and long-term consumption trends effectively. Overall, the peak demand analysis validates the practical reliability of the proposed approach for real-world deployment in healthcare energy management systems, particularly in scenarios requiring accurate forecasting under high-load operational stress.

British Journal of Computer, Networking and Information Technology
ISSN 2689-5315 • Volume 9, Issue 2, 2026 • pp. 127–148

Figure 6: Comparison of prediction errors under peak and normal load conditions.



Source: Generated by the author from experimental results.

Model Diagnostics and Residual Analysis

To further evaluate the robustness and statistical reliability of the proposed forecasting model, a comprehensive diagnostic analysis was conducted on the Random Forest predictions. Residuals were computed over the full test dataset as the difference between observed and predicted energy demand values to ensure complete statistical validity.

For visualization purposes, a stratified random sample of observations was used due to dataset size and computational constraints, while preserving the representativeness and statistical structure of the full dataset. This approach ensures clarity of visualization without compromising the integrity of the residual diagnostics.

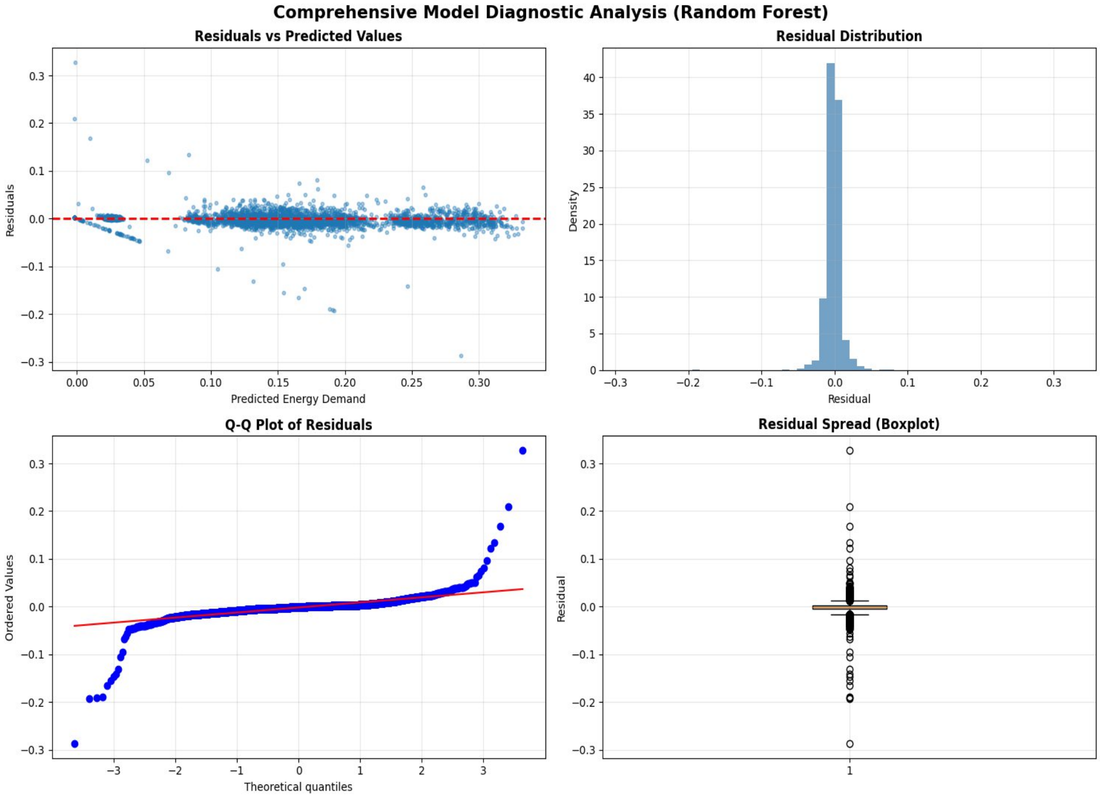
The residuals versus predicted values indicate that prediction errors are randomly distributed around zero, with no visible structural pattern, suggesting that the model effectively captures the nonlinear relationships underlying building energy consumption. The absence of curvature or funnel-shaped dispersion further confirms that no significant systematic bias or heteroscedasticity is present.

The residual distribution is sharply centered around zero and approximately symmetric, indicating near-Gaussian error behavior with minor tail deviations. The Q-Q plot further supports this observation, with most points closely following the reference line and slight deviations at the extremes, which are typical in real-world energy systems due to irregular load fluctuations.

Finally, the boxplot highlights a compact interquartile range with a limited number of outliers, demonstrating the stability of the model across varying operating conditions. Collectively, these results confirm that the proposed model produces statistically well-behaved, unbiased, and reliable predictions.

British Journal of Computer, Networking and Information Technology
ISSN 2689-5315 • Volume 9, Issue 2, 2026 • pp. 127–148

Figure 7: Comprehensive diagnostic analysis of the Random Forest model including residuals vs predicted values, residual distribution, Q-Q plot, and residual spread.



Source: Generated by the author from experimental results.

Practical Implications for Healthcare Energy Management

The findings of this study have important implications for real-world healthcare building energy management systems. The dominance of short-term autoregressive features (particularly lag_1) suggests that near-real-time monitoring of energy consumption can provide highly informative signals for short-horizon operational decision-making. This indicates that building energy management systems (BEMS) can achieve strong predictive performance even with limited external data inputs, provided that high-resolution consumption histories are available.

The strong influence of building size further implies that energy demand forecasting systems can benefit from facility-specific calibration, where baseline consumption profiles are adjusted according to structural characteristics such as floor area. This supports the development of scalable, building-aware forecasting systems that require minimal retraining across similar healthcare facilities.

From an operational perspective, the ability of ensemble models to accurately predict peak and non-peak demand enables improved load scheduling, demand response participation, and HVAC optimization. In particular, the reliable identification of short-term demand patterns can assist facility managers in preemptive load balancing and peak shaving strategies, thereby improving energy efficiency and reducing operational costs.

Finally, the limited contribution of external meteorological variables under the evaluated conditions suggests that, for short-term forecasting in healthcare environments, internal consumption dynamics may be prioritized over complex external data integration. This insight is valuable for practitioners seeking lightweight yet accurate forecasting solutions in operational building energy systems.

Limitations of the Study

Despite the strong predictive performance achieved, several limitations should be acknowledged. First, the proposed framework relies on historical energy and meteorological data and may therefore be less robust under non-stationary operating conditions or abrupt changes in building operation. Future work should investigate the integration of real-time occupancy information and adaptive learning strategies.

Second, the LSTM and GRU models were implemented using representative fixed architectures to provide practical deep learning baselines within a unified experimental framework. More comprehensive hyperparameter optimization or neural architecture search may further improve their performance. Consequently, the reported comparisons should be interpreted as an evaluation of practical forecasting pipelines rather than fully optimized implementations of each algorithm.

Third, the competitive performance of the naïve persistence model reflects the strong short-term temporal dependence present in healthcare energy demand and may not extend to building types with more volatile load characteristics.

Finally, although the ASHRAE Great Energy Predictor III dataset covers multiple climate zones, this study was limited to the healthcare subset comprising 23 buildings. Extending the proposed framework to additional building categories and larger cross-building evaluations would provide stronger evidence of its broader applicability.

CONCLUSION

This study developed and evaluated an explainable machine learning framework for healthcare building energy demand forecasting using ensemble learning, recurrent neural networks, and a naïve persistence baseline. Under the adopted experimental framework, Random Forest achieved the strongest predictive performance, followed closely by XGBoost, while the GRU and LSTM models produced comparatively higher prediction errors. These findings indicate that ensemble learning methods are well suited to structured, feature-rich healthcare energy datasets and provide an effective balance between predictive accuracy and model interpretability.

The inclusion of GRU, LSTM, and the naïve persistence model provides a comprehensive benchmark for evaluating forecasting approaches. Although the recurrent neural network models capture temporal dependencies, their performance under the adopted experimental configuration was lower than that of the ensemble models. The competitive SMAPE achieved by the naïve persistence model further highlights the strong short-term temporal dependence present in healthcare energy demand, suggesting that recent historical observations contain substantial predictive information for short-horizon forecasting.

Feature importance and SHAP-based interpretability analysis reveal that recent energy consumption (lag_1) and building size (square_feet) are the dominant drivers of energy demand prediction, while meteorological and calendar-based variables contribute marginally to the overall model behaviour. Ablation and cross-building validation analyses further provide complementary evidence on the role of engineered temporal features and support the robustness of the proposed framework under unseen-building conditions. Residual diagnostics indicate low prediction bias, near-zero mean error, and well-behaved error distributions, supporting the statistical consistency and reliability of the predictive models.

Overall, the findings suggest that explainable ensemble learning provides an accurate and interpretable approach for high-resolution healthcare building energy demand forecasting. The study further highlights the value of rigorous benchmarking through the inclusion of naïve persistence, deep learning, and ensemble learning models, thereby promoting more transparent and reliable evaluation of machine learning approaches for intelligent building energy management. The integration of explainable machine learning further enhances the practical value of the proposed framework by providing transparent model interpretation and creating opportunities for future integration with intelligent building monitoring and anomaly detection systems (Noorchenarboo and Grolinger 2025). Future research should investigate broader validation across additional building categories and more extensively optimized deep learning architectures to further assess the applicability of the proposed framework.

AI Declaration

The author used ChatGPT (OpenAI) to improve the clarity, grammar, organization, and readability of the manuscript. The author carefully reviewed, edited, and verified all AI-assisted content and takes full responsibility for the final content of the submitted work.

REFERENCES

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2015). Tree-based ensemble methods for predicting building energy consumption. *Renewable and Sustainable Energy Reviews, 47*, 213–227. https://doi.org/10.1016/j.rser.2015.03.065

Amiri, S. S., Mueller, M., & Hoque, S. (2023). Investigating the application of a commercial and residential energy consumption prediction model for urban planning scenarios with machine learning and Shapley additive explanation methods. *Energy and Buildings, 287*, Article 112965. https://doi.org/10.1016/j.enbuild.2023.112965

Bawaneh, K., Ghazi Nezami, F., Rasheduzzaman, M., & Deken, B. (2019). Energy consumption analysis and characterization of healthcare facilities in the United States. *Energies, 12*(19), Article 3775. https://doi.org/10.3390/en12193775

Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy, 9*, Article 100123. https://doi.org/10.1016/j.adapen.2023.100123

Cui, X., Lee, M., Koo, C., & Hong, T. (2024). Energy consumption prediction and household feature analysis for different residential building types using machine learning and SHAP: Toward energy-efficient buildings. *Energy and Buildings, 309*, Article 113997. https://doi.org/10.1016/j.enbuild.2024.113997

Devanathan, B., Jnana Varshitha, K., Pavan Kumar, L., Lakshmanan, S. A., & Krishna Prakash, N. (2025). Explainable AI framework using XGBoost with SHAP and LIME for multi-scale household energy forecasting. *IEEE Access, 13*, 149750–149764. https://doi.org/10.1109/ACCESS.2025.3602673

Dong, J., Zhou, J., Hamada, A., Paydar, M. A., Jain, N., Yang, D., Smith, N. M., & Burman, E. (2026). A data-driven approach for building energy efficiency of healthcare facilities in England and Wales using regulatory building energy datasets. *Building* *and Environment*. Advance online publication. https://doi.org/10.1016/j.buildenv.2026.114997

Fatehijananloo, M., Stopps, H., & McArthur, J. J. (2024). Exploring artificial intelligence methods for energy prediction in healthcare facilities: An in-depth extended systematic review. *Energy and Buildings, 320*, Article 114598. https://doi.org/10.1016/j.enbuild.2024.114598

González González, A., García-Sanz-Calcedo, J., & Rodríguez Salgado, D. (2018). Evaluation of energy consumption in German hospitals: Benchmarking in the public sector. *Energies, 11*(9), Article 2279. https://doi.org/10.3390/en11092279

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Liu, Q., Yu, D., Xue, Y., Sun, H., Yang, S., & Han, F. (2026). A cross-building few-shot energy load forecasting framework via transferable temporal learning and residual correction. *Energy and Buildings, 363*, Article 117530. https://doi.org/10.1016/j.enbuild.2026.117530

M.Sadeeq, M. A. (2025). XDL-energy: Explainable hybrid deep learning framework for smart campus energy forecasting and anomaly justification using the DPU-ALDOSKI dataset. *Energy and Buildings, 347*, Article 116305. https://doi.org/10.1016/j.enbuild.2025.116305

Ma, Z., Jiang, G., & Chen, J. (2024). Physics-informed ensemble learning with residual modeling for enhanced building energy prediction. *Energy and Buildings, 323*, Article 114853. https://doi.org/10.1016/j.enbuild.2024.114853

Marzouk, O. A. (2025). Summary of the 2023 (1st edition) report of TCEP (Tracking Clean Energy Progress) by the International Energy Agency (IEA), and proposed process for computing a single aggregate rating. *E3S Web of Conferences, 601*, Article 00048. https://doi.org/10.1051/e3sconf/202560100048

Miller, C., Arjunan, P., Kathirgamanathan, A., Fu, C., Roth, J., Park, J. Y., Balbach, C., Gowri, K., Nagy, Z., Fontanini, A. D., & Haberl, J. (2020). The ASHRAE Great Energy Predictor III competition: Overview and results. *Science and Technology for the Built Environment, 26*(10), 1427–1447. https://doi.org/10.1080/23744731.2020.1795514

Miller, C., Hao, L., & Fu, C. (2022). Gradient boosting machines and careful pre-processing work best: ASHRAE Great Energy Predictor III lessons learned. *ASHRAE Transactions, 128*(2), 405–413.

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*, Article 368. https://doi.org/10.1038/s41597-020-00712-x

Neubauer, A., Brandt, S., & Kriegel, M. (2025). Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability. *Energy and AI, 20*, Article 100480. https://doi.org/10.1016/j.egyai.2025.100480

Noorchenarboo, M., & Grolinger, K. (2025). Explaining deep learning-based anomaly detection in energy consumption data by focusing on contextually relevant data. *Energy and Buildings, 328*, Article 115177. https://doi.org/10.1016/j.enbuild.2024.115177

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings, 40*(3), 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Somu, N., Raman M. R., G., & Ramamritham, K. (2021). A deep learning framework for building energy consumption forecast. *Renewable and Sustainable Energy Reviews, 137*, Article 110591. https://doi.org/10.1016/j.rser.2020.110591

Wang, M., Zheng, Y., Wang, B., & Deng, Z. (2021). Household electricity load forecasting based on multitask convolutional neural network with profile encoding. *Mathematical Problems in Engineering, 2021*, Article 6661798. https://doi.org/10.1155/2021/6661798

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., & Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. *Energy and Buildings, 171*, 11–25. https://doi.org/10.1016/j.enbuild.2018.04.008

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews, 82*, 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108

Zhang, L., & Chen, Z. (2024). Large language model-based interpretable machine learning control in building energy systems. *Energy and Buildings, 313*, Article 114278. https://doi.org/10.1016/j.enbuild.2024.114278

Zhao, A., Chen, M., Quan, W., & Zhang, S. (2024). A hybrid forecasting model for general hospital electricity consumption based on mixed signal decomposition. *Energy and Buildings, 325*, Article 115006. https://doi.org/10.1016/j.enbuild.2024.115006

Zhao, H.-X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews, 16*(6), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049

Abbreviations

Abbreviation	**Full meaning**
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CNN	Convolutional Neural Network
CNN–LSTM	Convolutional Neural Network–Long Short-Term Memory
GEPIII	Great Energy Predictor III
GRU	Gated Recurrent Unit
HVAC	Heating, Ventilation, and Air Conditioning
LIME	Local Interpretable Model-Agnostic Explanations
LightGBM	Light Gradient Boosting Machine

Abbreviation	Full meaning
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
NHS	National Health Service
RF	Random Forest
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
SMAPE	Symmetric Mean Absolute Percentage Error
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting