Hindawi
International Journal of Energy Research
Volume 2024, Article ID 6812425, 19 pages
https://doi.org/10.1155/2024/6812425

WILEY | Hindawi

Research Article

Machine Learning Algorithms for Predicting Energy Consumption in Educational Buildings

Khaoula Elhabyb,[1] **Amine Baina**,[1] **Mostafa Bellafkih**,[1] **and Ahmed Farouk Deifalla**[2]

[1]*National Institute of Posts and Telecommunications (INPT), Rabat, Morocco*
[2]*Future University Cairo in Egypt, Cairo, Egypt*

Correspondence should be addressed to Ahmed Farouk Deifalla; ahmed.deifalla@fue.edu.eg

Received 14 December 2023; Revised 19 March 2024; Accepted 26 March 2024; Published 13 May 2024

Academic Editor: Saleh N. Al-Saadi

In the past few years, there has been a notable interest in the application of machine learning methods to enhance energy efficiency in the smart building industry. The paper discusses the use of machine learning in smart buildings to improve energy efficiency by analyzing data on energy usage, occupancy patterns, and environmental conditions. The study focuses on implementing and evaluating energy consumption prediction models using algorithms like long short-term memory (LSTM), random forest, and gradient boosting regressor. Real-life case studies on educational buildings are conducted to assess the practical applicability of these models. The data is rigorously analyzed and preprocessed, and performance metrics such as root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used to compare the effectiveness of the algorithms. The results highlight the importance of tailoring predictive models to the specific characteristics of each building's energy consumption.

1. Introduction

Artificial intelligence is rapidly being integrated into various industries, such as healthcare, finance, and smart grids. Among these human-centric applications, the use of AI in smart buildings has attracted significant attention from a large community [1]. Smart buildings, which have been a subject of research since the 1980s, utilize advanced technology, data analytics, and automation systems to optimize operations, enhance occupant comfort and productivity, and reduce costs and energy consumption [2]. These buildings incorporate sensors, devices, and control systems to monitor lighting, HVAC systems, security, and access controls. Real-time data on occupancy, temperature, air quality, and energy use can be analyzed to identify optimization opportunities. The primary aim is to create an efficient, comfortable, and sustainable environment for residents while reducing costs and ecological impact.

The smart building industry is experiencing significant growth as society becomes more connected and digital. According to statistics from MarketsandMarkets [3], the industry is projected to expand at a compound annual growth rate (CAGR) of 10.5% between 2020 and 2025, reaching a value of $108.9 billion. This growth is driven by factors such as increased energy usage and expenses, advancements in machine learning and the Internet of Things (IoT), the push for net zero energy buildings, and regulatory changes that encourage the adoption of smart building systems and services. Figure 1 presents the forecasted global market side from 2020 to 2030. Expanding on the findings of the Zion Marketing research study [4], it reveals the market value of 40,760 million in 2016, with projections of a substantial growth trajectory to 61,900 million by 2024, with a CAGR exceeding 34%. This indicates a rapid expansion within the market, indicating robust trends and significant economic development during the study period.

The AI sector being discussed is experiencing significant growth due to the integration of the Internet of Things (IoT) and machine learning (ML). IoT sensors collect data about buildings and occupants, such as temperature, humidity, occupancy, and electricity consumption.

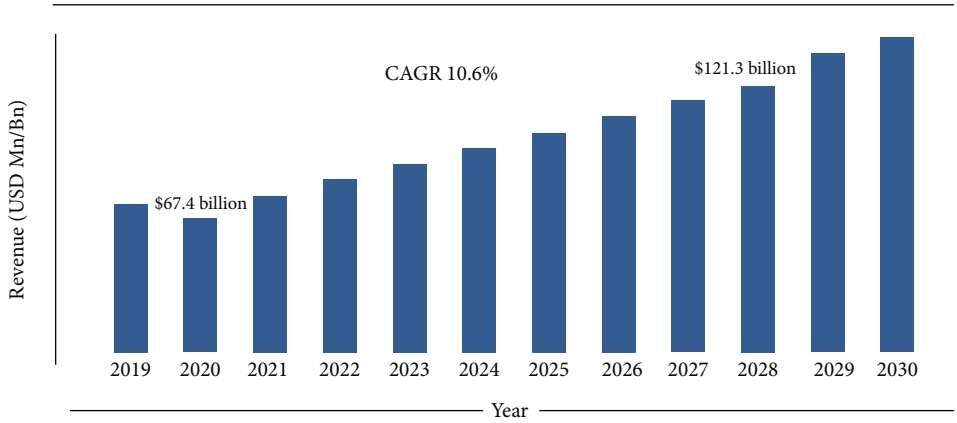

Figure 1: The global smart building market size [3].

This data is centralized for optimizing building operations, improving resident comfort, and reducing energy usage. ML, on the other hand, is a powerful tool for processing large amounts of data from various sources. It analyzes this data to identify patterns and predict future events, such as equipment failures, enabling preventative maintenance [5].

The American Council for Energy-Efficient Economy (ACEEE) [6] suggests that commercial buildings can significantly reduce their energy bills by up to 30% by implementing energy-efficient technologies such as smart thermostats and controlled lighting. The US Department of Energy [7] reports that commercial buildings account for a significant portion of total energy consumption and greenhouse gas emissions in the US. This highlights the importance of buildings that can predict energy consumption and plan efficiently to reduce energy usage. Intel research [8] indicates also that energy consumption prediction has the potential to achieve operational cost savings, staff productivity gains, and energy usage reductions. Given these findings, the primary emphasis will be on forecasting the energy usage of smart buildings, with a specific focus on educational facilities, which will be analyzed for the first time. Understanding and predicting energy consumption in educational environments are paramount for optimizing resource allocation, implementing effective efficiency measures, and establishing sustainable and cost-effective operational procedures [9]. By focusing on this sector, valuable insights can be gained to inform strategies for enhancing energy efficiency and sustainability in educational buildings, ultimately contributing to improved resource management and environmental conservation efforts.

The research concentrates on energy management within smart buildings, aiming to forecast power consumption through three distinct approaches: a traditional statistical approach employing the random forest algorithm, a deep learning approach utilizing long short-term memory (LSTM), and a hybrid approach leveraging the gradient boosting regressor algorithm. These three techniques were chosen to investigate a research gap regarding to the majority of data-driven methodologies. While significant progress has been made in this area, limited attention has been given to utilizing streaming and temporal data for forecasting buildings' energy demand. This gap will be addressed through the utilization of real historical electricity data. The used data is analyzed to evaluate model performance and accuracy, aiming to identify the most effective approach for smart building energy management. The research is aimed at optimizing forecasting techniques through rigorous comparative analysis, leveraging the strengths of LSTM, RF, and GBR models. The study highlights the importance of advanced machine learning in shaping smart building strategies and is aimed at enhancing sustainability and efficiency in energy usage. Insights from this research will inform future advancements in energy management practices for sustainable development. The article is structured into several delineated sections, each serving a specific purpose:

(i) Introduction: This section introduces the application of AI within the smart building sector, setting the context for the study

(ii) Literature analysis: Here, a comparative examination of various ML algorithms used for energy prediction in smart building systems is provided, drawing insights from existing research

(iii) Methodology: This section outlines the systematic approach adopted in the study, encompassing data analytics, model development, and model evaluation processes

(iv) Results and discussions: Findings obtained from the methodology are presented, followed by a comparative analysis that juxtaposes these results with prior research initiatives

(v) Conclusion: This section synthesizes the results and provides conclusions, offering perspectives on the implications of the study's findings for the field of smart building energy management

2. Literature Review

A recent study conducted by the International Energy Agency [10] has revealed concerning levels of energy

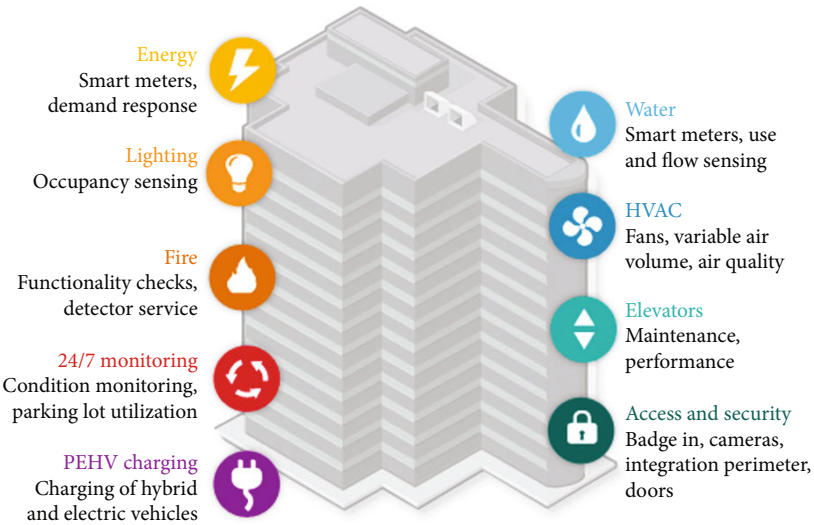

Figure 2: The global smart building market size [11].

consumption in buildings. The study found that buildings are responsible for a significant portion of electricity consumption and overall energy consumption in urban areas. Buildings account for 72% of total electricity consumption and 38% of average energy consumption in urban areas. Additionally, buildings contribute to almost 40% of total carbon dioxide pollution in urban areas. A smart building is a modern infrastructure that incorporates automated control systems and uses data to improve the building's performance and occupants' comfort. Figure 2 presents the smart building functionalities and its most important axis of work.

The top technology companies are currently prioritizing IoT (Internet of Things) and AI (artificial intelligence). The future of building innovation is expected to focus on achieving maximum energy efficiency, and this challenge can be addressed by integrating AI-powered systems like machine learning (ML) and deep learning. ML systems continuously improve themselves, leading to advancements in various AI research areas [12]. ML involves algorithms that allow them to respond to inputs from their environment and identify nonlinear connections in complicated or uncertain systems. ML is divided into four major categories based on the type of learning task they manage: supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning.

(i) Supervised learning is a method of developing a machine learning model by using a labeled data set. In this process, each data point in the set is associated with a known intended output. The model is trained to predict the output

(ii) Unsupervised learning: in contrast to traditional supervised learning, developing a model on an unlabeled data set involves working with data where the target outputs are unknown. In this scenario, the model is not explicitly instructed on what to search for but instead learns

(iii) Semisupervised learning is a learning approach that combines supervised and unsupervised learning. In this approach, the model is trained using a data set that is partly labeled, meaning that some of the data points have known labels

(iv) Reinforcement learning where a model is trained to make a series of decisions in a changing environment. The model learns through trial and error, receiving feedback in the form of rewards or costs

Energy consumption prediction is a valuable technique that involves forecasting the amount of energy a system or device will use within a specific time frame. This technique serves various purposes, such as optimizing energy usage, predicting future energy demands, and identifying potential inefficiencies in energy consumption. To predict energy consumption, different methods can be employed, including statistical models, machine learning algorithms, and physics-based models. The choice of technique depends on factors such as data availability, system complexity, and the desired level of accuracy. In this particular case, the focus is on utilizing machine learning algorithms to predict energy consumption by leveraging historical data and other relevant factors.

The quality and relevance of the data used in machine learning algorithms greatly influence their performance. In a study conducted by Ahajjam et al. [13] on Moroccan Buildings' Electricity Consumption Data Set, electricity consumption was categorized into three types: whole premises (WP), individual loads (IL), and circuit-level (CL) data.

(1) Labeled WP: Labeled whole premises (WP) consumption data refers to electricity usage data collected from 13 households in the MORED data set.

This data is valuable as it includes not only the raw electricity consumption measurements but also additional information that can assist in analyzing, modeling, and comprehending the patterns of electricity usage in different households

(2) Labeled IL: Ground-truth electricity consumption refers to the electricity consumption data of individual loads (IL) that have been labeled or annotated with accurate information. This involves recording and labeling the operational states of specific loads, such as refrigerators or air conditioners when they are turned on or off at specific times. Having this ground-truth information is valuable for researchers and analysts as it allows for accurate load disaggregation, energy management, and appliance recognition

(3) CL: Measurements in the context of energy refer to the circuit-level energy measurements obtained from the electrical mains of a premises. These measurements provide information about the overall energy consumption of a circuit and can be used to understand the energy consumption of a group of loads

The current work focuses on three educational buildings located at Down Town University. Further information about these buildings will be provided next. The subsequent section presents a literature review on energy consumption forecasting in various buildings using multiple machine learning algorithms.

2.1. Traditional Machine Learning Approach. ML algorithms have been utilized to tackle the primary challenges of physics-driven methods in load prediction. For instance, Somu et al. [14] developed eDemand, a new building energy use forecasting model, using long short-term memory networks and an improved sine cosine optimization algorithm, and as a result, the model outperformed previous state-of-the-art models in real-time energy load prediction. Next, Suranata et al. [15] focused on predicting energy consumption in kitchens. They used a feature engineering technique and a short-term memory (LSTM) model. Principal component analysis (PCA) was applied to extract important features, and the LSTM model was used on two tables. In addition, Shapi et al. [16] developed a prediction model for energy demand making use of the Microsoft Azure cloud-based machine learning framework, The methodology of the prediction model is provided using three distinct techniques, including support vector machine, artificial neural network, and k-nearest neighbors. The study focuses on real-world applications in Malaysia, with two tenants from an industrial structure chosen as case studies. The experimental findings show that each tenant's energy consumption has a particular distribution pattern, and the suggested model can accurately estimate energy consumption for each renter. To forecast daily energy consumption based on weather data, Faiq et al. [17] developed a new energy usage prediction technique for institutional buildings using long short-term memory (LSTM). The model, trained using Malaysian Meteorological Department weather forecasting data, outperformed support vector regression (SVR) and Gaussian process regression (GPR) with the best RMSE scores. The dropout method reduces overfitting, and Shapley's additive explanation is used for feature analysis. Accurate energy consumption estimates can help detect and diagnose system faults in buildings, aiding in energy policy implementation. Further, Kawahara et al. [18] explore the application of various machine learning models to predict voltage in lithium-ion batteries. The study includes algorithms such as support vector regression, Gaussian process regression, and multilayer perceptron. The hyperparameters of each model were optimized using 5-fold cross-validation on training data. The data set used consists of both simulation data, generated by combining driving patterns and applying an electrochemical model, and experimental data. The performance of the ML models was evaluated using both simulation and experimental data, with different data sets created to simulate variations in state of charge distribution.

2.2. Deep Learning and Hybrid Approaches. Additionally, various networks integrate multiple techniques to devise data-driven approaches. These integrated mechanisms are commonly referred to as hybrid networks. For example, Mohammed et al. [19] focus on the application of an intelligent control algorithm in HVAC systems to enhance energy efficiency and thermal comfort. The authors propose integrating SCADA systems with an intelligent building management system to optimize heat transmission coefficients and air temperature values. Genetic algorithms are employed to maintain user comfort while minimizing energy consumption. Similar to [19], Aurna et al. [20] compare the performance of ARIMA and Holt-Winters models in predicting energy consumption data in Ohio and Kentucky. The study finds that the Holt-Winters model is more accurate and effective for long-term forecasting. The authors recommend further research to consider other parameters, and environmental factors, and explore hybrid models for better short-term load forecasting. Next, Ferdoush et al. [21] developed a hybrid forecasting model for time series electrical load data. The model combines random forest and bidirectional long short-term memory methods and was tested on a 36-month Bangladeshi electricity consumption data set. The results showed that the hybrid model outperformed standard models in terms of accuracy. The study emphasizes the effectiveness of the hybrid machine learning approach in improving short-term load forecasting accuracy in the dynamic electric industry. In their study, He and Tsang [22] developed a hybrid network combining long short-term memory (LSTM) and improved complete ensemble empirical mode decomposition with adaptive noise (iCEEMDAN) to optimize electricity consumption. They divided the initial power consumption data into patterns using iCEEMDAN and used Bayesian-optimized LSTM to forecast each mode independently. In the same direction, Jin et al. [23] proposed an attention-based encoder-decoder network with Bayesian optimization for short-term electrical load forecasting, using a gated recurrent unit recurrent neural network for time series data modeling and a temporal attention layer for improved prediction accuracy and

precision. Further in their study, Olu-Ajayi et al. [24] used various machine learning techniques to predict yearly building energy consumption using a large data set of residential buildings. The model allows designers to enter key building design features and anticipate energy usage early in the development process. DNN was found to be the most efficient predictive model, motivating building designers to make informed choices and optimize structures. Jang et al. [25] created three LSTM models to compare the effects of incorporating operation pattern data on prediction performance. The model using operation pattern data performed the best, with a CVRMSE of 17.6% and an MBE of 0.6%. The article by Ndife et al. [26] presents a smart power consumption forecast model for low-powered devices. The model utilizes advanced methodologies, such as the ConvLSTM encoder-decoder algorithm, to accurately predict power consumption trends. The performance evaluation of the model demonstrates improved accuracy and computational efficiency compared to traditional methods. Also, Duong and Nam [27] developed a machine learning system that monitors electrical appliances to improve electricity usage behavior and reduce environmental impact. The system utilizes load and activity sensors to track energy consumption and operating status. After three weeks of testing, the system achieved a state prediction accuracy of 93.60%. In their approach, Vennila et al. [28] propose a hybrid model that integrates machine learning and statistical techniques to improve the accuracy of predicting solar energy production. The model also helps in reducing placement costs by emphasizing the significance of feature selection in forecasting. In the sale context, Kapp et al. [29] developed a supervised machine learning model to address energy use reduction in the industrial sector. They collected data from 45 manufacturing sites through energy audits and used various characteristics and parameters to predict weather dependency and production reliance. The results showed that a linear regressor over a transformed feature space was a better predictor than a support vector machine. In their research, Bhol et al. [30] propose a new method for predicting reactive power based on real power demand. They utilize a flower pollination algorithm to optimize their model and show that it outperforms other models like GA, PSO, and FPA. Asiri et al. [31] used an advanced deep learning model for accurate load forecasting in smart grid systems. They use hybrid techniques, including LSTM and CNN, feature engineering, and wavelet transforms, to enhance forecasting accuracy and efficiency. The results show significant improvements in short-term load prediction, outperforming traditional forecasting methods.

Table 1 contains detailed information about the algorithms used, performance evaluation measurements, and the advantages and disadvantages of each approach.

3. Methodology

This research predicts power usage in three buildings of a private research university using a data set collected from January 2020 to January 2023. The university is known for its excellence in education and research across various disciplines. The buildings under study (referred to as CLAS, NHAI, and Cronkite) are all part of the same institution and serve distinct functions. Building CLAS, an abbreviation of Center of Law and Society, mainly consists of an amphitheater and offices, and building NHAI, which means Nursing and Health Innovation, consists of offices and laboratories. In contrast, Cronkite consists of classrooms and seminar halls.

The buildings are equipped with IoT sensors connected to power intel sockets, and the collected data is sorted on an open-source website server [32]. The prediction method will use three machine learning algorithms: long short-term memory (LSTM), random forest (RF), and gradient boosting regressor (GBR). The data will be analyzed and prepared before being used to train and test the models.

The methodology for forecasting energy consumption will be divided into three sections:

(1) Data analysis involves evaluating raw data to understand patterns and characteristics of electrical power consumption data.

(2) Model training trains machine learning models, using past data to identify patterns and correlations between input characteristics and day power use

(3) Model test models evaluation using validation metrics to assess their performance and accuracy.

3.1. Data Analysis

3.1.1. Data Preparation. This study focuses on the process of data preparation in machine learning, which is time-consuming and computationally challenging due to the presence of missing values and uneven value scales between features. The data was prepared using two techniques: imputation of missing data and standardization. The imputation procedure was carried out using the probabilistic principal component analysis (PPCA) approach, a maximum likelihood estimate-based technique that estimates missing values using the expectation-maximization (EM) algorithm. This method is developed from the principal component analysis (PCA) method, which is used for data compression or dimensionality reduction. The resulting cleaned data was then subjected to standardization, also known as Z-score normalization, to ensure an even distribution of the data above and below the mean value as shown in equation (3):

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ represents the mean, σ denotes the standard deviation, and x is the original data points.

3.1.2. Data Normality Analysis. This research conducted a normality test on each renter's data set to determine its distribution. This test is crucial for model construction and is especially important for larger sample sizes. Understanding the data set distribution can provide valuable insights into the prediction outcome. Kurtosis measures distribution peaks,

Table 1: Previous research in ML-driven building energy use prediction.

Authors	Algorithm	Data set	Performance evaluation	Pros	Cons
Somu et al. [14]	(i) ARIMA (ii) Genetic algorithm-LSTM (iii) Sine cosine optimization algorithm-LSTM	(i) The KReSIT power consumption data set is sourced from the Indian Institute of Technology (IIT) in Mumbai, India.	(i) ARIMA: MAE: 0.3479; MAPE: 21.3333; MSE: 0.1661; RMSE: 0.4076. (ii) Genetic algorithm-LSTM: MAE: 0.1804; MAPE: 5.9745; MSE: 0.0432; RMSE: 0.2073. (iii) (ISCOA-LSTM): MAE: 0.0819; MAPE: 4.9688; MSE: 0.0135; RMSE: 0.1164.	(i) Improved forecasting accuracy (ii) Improved forecasting accuracy (iii) Real-world applicability	(i) Sensitivity to initialization (ii) Convergence speed
Suranata et al. [15]	(i) Long short-term memory	(i) NL	(i) RMSE = 62.013; MAE = 26.982; MAPE = 12.876	(i) The ability to effectively predict energy consumption patterns in time series data.	(i) Time-consuming training
Ferdoush et al. [21]	(i) LSTM (ii) RF-bi-LSTM hybrid model (iii) Bidirectional long short-term memory (LSTM)	(i) Bangladesh Power Development Board covered 36 months.	(i) LSTM: MSE = 0.4776; RMSE = 0.691; MAE = 0.5578; MAPE = 148.7. (ii) Bi-LSTM: MSE = 0.2943; RMSE = 0.5425; MAE = 0.4317; MAPE = 194.80. (iii) RF-bi-LSTM: MSE = 0.1673; RMSE = 0.4090; MAE = 0.3070 MAPE = 193.49.	(i) Stable learning characteristics. (ii) Moderate generalization gap in learning loss analysis.	(i) The hybrid model may require specific data to utilize the strengths of random forest and bidirectional LSTM effectively.
Yaqing et al. [22]	(i) EMD-BO-LSTM (ii) iCEEMDAN	(i) Real power consumption data of a university campus for 12 months	(i) EMD-BO-LSTM: MAE = 155.77; RMSE = 203.4; MAPE = 10.41%; R2 = 0.8478. (ii) iCEEMDAN-BO-LSTM: MAE = 40.841; RMSE = 59.68; MAPE = 2.5563%; R2 = 0.986.	(i) Adaptability and efficiency (ii) Enhanced prediction accuracy	(i) NL
Ndife et al. [26]	(i) ConvLSTM encoder-decoder	(i) Two million measurements were gathered over 47 months from a residential location in Sceaux, France.	(i) RMSE on the model: 358 kWh RMSE on the persistence model: 465 kWh RMSE on model A: 530 kWh RMSE on model B: 450.5 kwh	(i) Improved forecast accuracy (ii) Suitable for low-powered devices (iii) Efficient training and prediction time	(i) Model complexity
Duong et al. [27]	(i) Multiple layer perceptron	(i) 215 data points on the power consumption and on/off status of electrical devices, in Vietnam.	(i) RMSE: 10.468 (ii) MAPE: 21.563	(i) It handles large amounts of input data well. Makes quick predictions after training.	(i) Slow training
Faiq et al. [17]	(i) LSTM (ii) LSTM-RNN (iii) CNN-LSTM	(i) Daily data from 2018 to 2021, from the Malaysian Meteorological Department.	(i) LSTM: RMSE = 165.20; MAE = 572.545. (ii) LSTM-RNN: RMSE = 263.14; MAE = 353.38. (iii) CNN-LSTM: RMSE = 692.14, MSE = 1134.1.	(i) Accurate prediction of building energy consumption (ii) Improved energy efficiency	(i) Requires a significant amount of historical data to create an accurate model
Bhol et al. [29]	(i) ARIMA (ii) Holt-Winters flower pollination algorithm	(i) Laboratory-operated critical loads over three months.	(i) HW-GFPA: MBE = 0.42 for validation, 0.43 for test RMSE = 0.80 (ii) ARIMA: MBE = 0.073 for validation, 0.016 for test RMSE = 0.183	(i) Scalability (ii) Optimal hyperparameter identification	(i) Sensitivity to kernel selection

while skewness measures irregular probability distribution around the mean value [33]. Equations (2) and (3) provide formulas for skewness and kurtosis, which are essential for understanding the data set distribution and its impact on the prediction outcome.

$$\text{Skewness} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N - 1 * \sigma^3}, \tag{2}$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{\sigma^4}, \tag{3}$$

where n is the number of data points in the collection, x_i is the individual data points within the sample, and $\bar{x}$ is the sample mean.

3.1.3. Feature Selection. Feature engineering is a crucial aspect of machine learning, involving the creation of meaningful data representations to enhance model performance. It involves careful selection, transformation, and creation of features that capture relevant information from raw data, enhancing predictive accuracy and interoperability. Techniques like principal component analysis, domain knowledge extraction, and creative data manipulation help models extract patterns and make accurate predictions, bridging the gap between raw data and actionable insights.

As previously stated, our data set comprises 27 features detailing the characteristics of the selected buildings. To ensure optimal input for our predictive model, we employed a feature engineering approach leveraging a tree-based model, specifically the random forest algorithm.

3.2. Model Development. This study uses supervised machine learning to predict energy usage using data prepared and trained in two groups. The model employs regressive prediction using random forest, LSTM, and gradient boosting regressor. The process from data collection to model generation is depicted in Figure 3.

3.2.1. Random Forest. A random forest regressor is a machine learning method that combines multiple decision trees to create a predictive model for regression tasks. Each tree is constructed using a randomly selected subset of training data and features with $H(x;\theta_k), k = 1, \cdots, K$ where x represents the observed input (covariate) vector of length p with associated random vector X. During prediction, the regressor aggregates predictions from all trees to generate the final output, typically the average of the individual three prediction $h(x) = (1/k)\sum_{k=1}^{K} h(x;\theta_k)$ [34]. This method is commonly used for pattern identification and prediction due to its ability to learn complicated behavior, Consequently, it is the best choice for constructing the prediction model in the present study. In Figure 4, we present a flow chart of the random forest algorithm.

3.2.2. Long Short-Term Memory. Sepp Hochreiter and Juergen Schmidhuber introduced long short-term memory (LSTM) in 1997 as an advanced application of recurrent neural networks. LSTM is effective in processing and predicting time series data with varying durations. It captures long-term relationships, handles variable-length sequences, and recalls previous data, making it useful for energy consumption prediction [35]. The LSTM model structure consists of three layers: input, LSTM unit, and output. The mathematical equations used in LSTM include the forget gate, input gate, output gate, and cell state. The following are the equations utilized in LSTM:

$$\begin{aligned} &i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \\ &C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \\ &o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) h_t = o_t \cdot \tan h(C_t), \end{aligned} \tag{4}$$

where x_t is the input at the step t; i_t, f_t, and o_t are the input, forgot, and output vectors; g_t is the candidate activation vector, and c_t is the cell state at time t.

The LSTM algorithm is a powerful tool for collecting and transmitting information across long sequences. It is commonly used in applications such as audio recognition, natural language processing, and time series analysis. Based on previous research and the availability of a time series data set, LSTM is chosen as the algorithm for predicting energy with high precision. Figure 5 presents a flowchart of LSTM.

3.2.3. Gradient Boosting Regressor. The gradient boosting approach is an iterative method that combines weak learners to create a strong learner by focusing on errors at every step. It is aimed at decreasing the loss function by finding an approximation function of the function $F(x)$ that translates x to y. This method improves prediction performance and lowers prediction error by matching weak learner models to the loss function [36]. The squared error function is often used to estimate the approximation function, which is then used to find the ideal settings for weak learners. The gradient boosting regressor's mathematical equation is as follows:

$$y_i = F(x_i) + \sum_{m=1}^{M} \gamma_m . h_{m(x_i)}, \tag{5}$$

where y_i is the predicted target, x_i is the input features, $F(x_i)$ is the ensemble model prediction, M is the weak model, γ_m is the learning rate, and $h_m(x_i)$ is the prediction by $m-$th weak model. The current research utilized gradient boosting due to its robust predictive performance, ability to capture complex data linkages and nonlinear patterns, and flexibility and customization capabilities. Figure 6 depicts the gradient boost regressor algorithm's flow chart.

3.3. Model Evaluation. The data set was divided into a training group (25%) and a testing group (75%). The training group was used to train machine learning algorithms and create predictive models for maximum consumption data. The testing group was used to evaluate the performance of these models. This process is illustrated in Figure 7.

The training and testing process involved a simple partitioning of data to prevent overfitting. Machine learning algorithms' predictive models were evaluated for performance

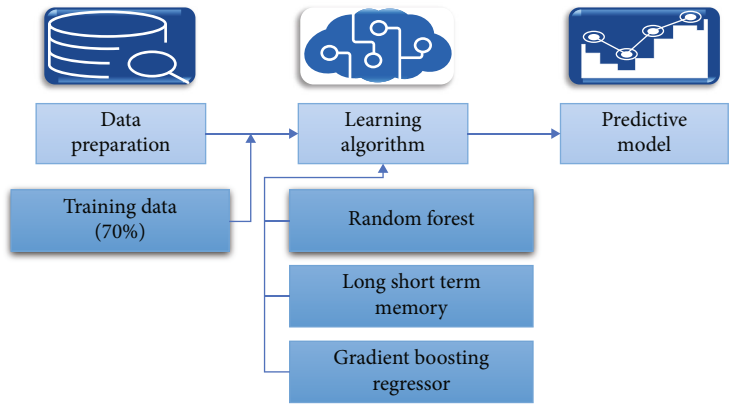

Figure 3: Process of generating predictive model after data preparation.

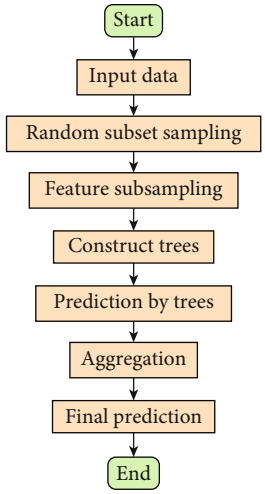

Figure 4: Random forest algorithm flowchart.

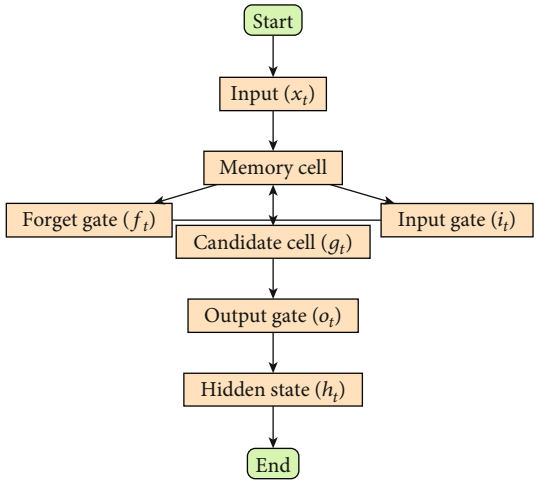

Figure 5: Long short-term memory algorithm flowchart.

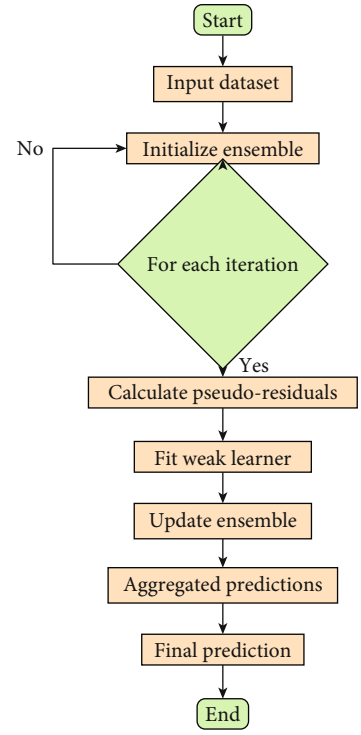

Figure 6: Gradient boosting regressor flow chart.

and accuracy using metrics like R2, MSE, MAE, RMSE, and MAPE. Each measurement definition is mentioned in Table 2.

The present research used MSE because of its sensitivity to errors, differentiability, and simplicity of interpretation. The use of RMSE is preferable to MSE because it yields a more easily understandable outcome in the original units of the dependent variable, facilitating straightforward comparison across data sets or models. The mean absolute error (MAE) is a suitable metric where the quantity of errors is more significant than the specific direction of the mistakes, offering a clear and direct evaluation of the model's performance, and MAPE is particularly valuable for comparing a model's prediction accuracy to the scale of the actual values.

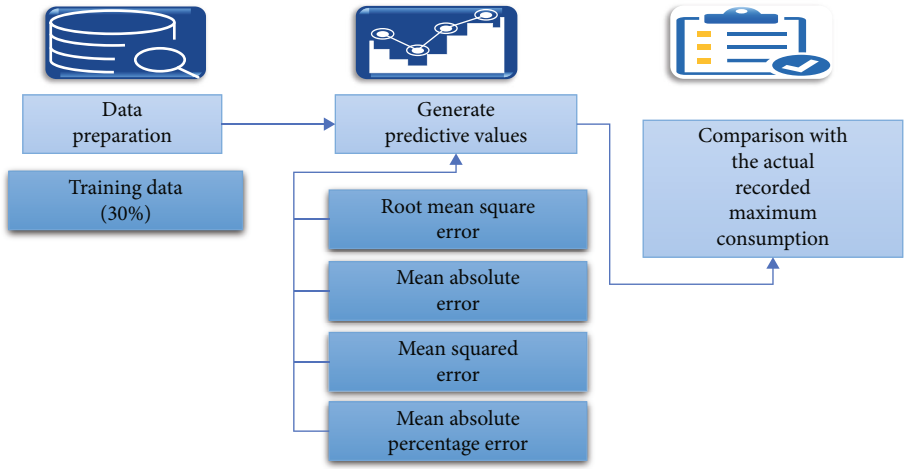

Figure 7: Testing procedure for the trained predictive model.

Table 2: Performance metrics.

Algorithms	Description	Math form
R-squared [37]	The coefficient of determination is used to determine how much of the variance in the dependent variable can be explained by the independent variables.	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
Mean squared error [38]	A regression metric used to calculate the average squared difference between predicted and actual values.	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Root mean squared error [39]	RMSE is a widely used measure for estimating the average variance between predicted and real values in regression tasks.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean absolute error [40]	A regression statistic used to calculate the average absolute difference between predicted and actual values, ignoring the direction of mistakes.	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Mean absolute percentage error [39]	A commonly used method for determining forecasting error, as it measures the average absolute percent inaccuracy for each time period less actual values divided by actual values, making understanding it simpler due to its scaled units.	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100$

4. Results and Discussion

The experiment results were reviewed in sections, discussing the initial processing and imputation of missing data, energy consumption prediction for each building, and performance comparisons for random forest, long short-term memory, and gradient boosting regressor models. The presentation of results follows a hierarchy, starting with the normality test, then data preprocessing, and finally model evaluation.

4.1. Normality Testing of Data. The evaluation is aimed at examining the impact of data shape on predictive model development performance, using measures of skewness and kurtosis. Results were compiled in Table 3 to evaluate the data's shape and potential deviations from normal distribution. To evaluate the normality of the energy demand data, the two values were computed using the aggregated data from each building spanning from January 2020 to January 2023. Figure 8 also depicts the format of the data set for a graphical examination of normality.

Based on Table 3, the data sets for the CLAS, NHAI, and Cronkite buildings were approximately symmetrical and skewed with bidirectional shape distribution. However, there were some differences in the skewness values for each building. The CLAS building showed normal asymmetry due to power consumption and KWS, with a slightly negative skewness indicating a longer left tail. The CHWTON distribution was skewed, with a skewness of 427578, indicating a longer left tail. The nursing and health innovation building had a pronounced asymmetry, with power consumption having a positive skewness and KWS and CHWTON having a negative skewness, indicating balanced tails. The Cronkite building had positive skewness values, indicating a moderate right-skewed distribution. Overall, all three data sets were approximately symmetric, skewed, and bimodal in their form density.

The kurtosis values of all three buildings in Table 3 were less than 0, indicating that their distributions were

Table 3: Measurements of skewness and kurtosis for the buildings.

Application area	Building name	Power consumption	KWS	CHWTON
Skewness	CLAS	-0.10206	-1	0.42329
	NHAI	0.017914	-1	-1
	Cronkite	0.805914	-1	0.494
Kurtosis	CLAS	-0.333548	-2.0	-1.019492
	NHAI	-0.777519	-2.0	-2
	CRONKIT	2.620576	2.056333	-0.687182

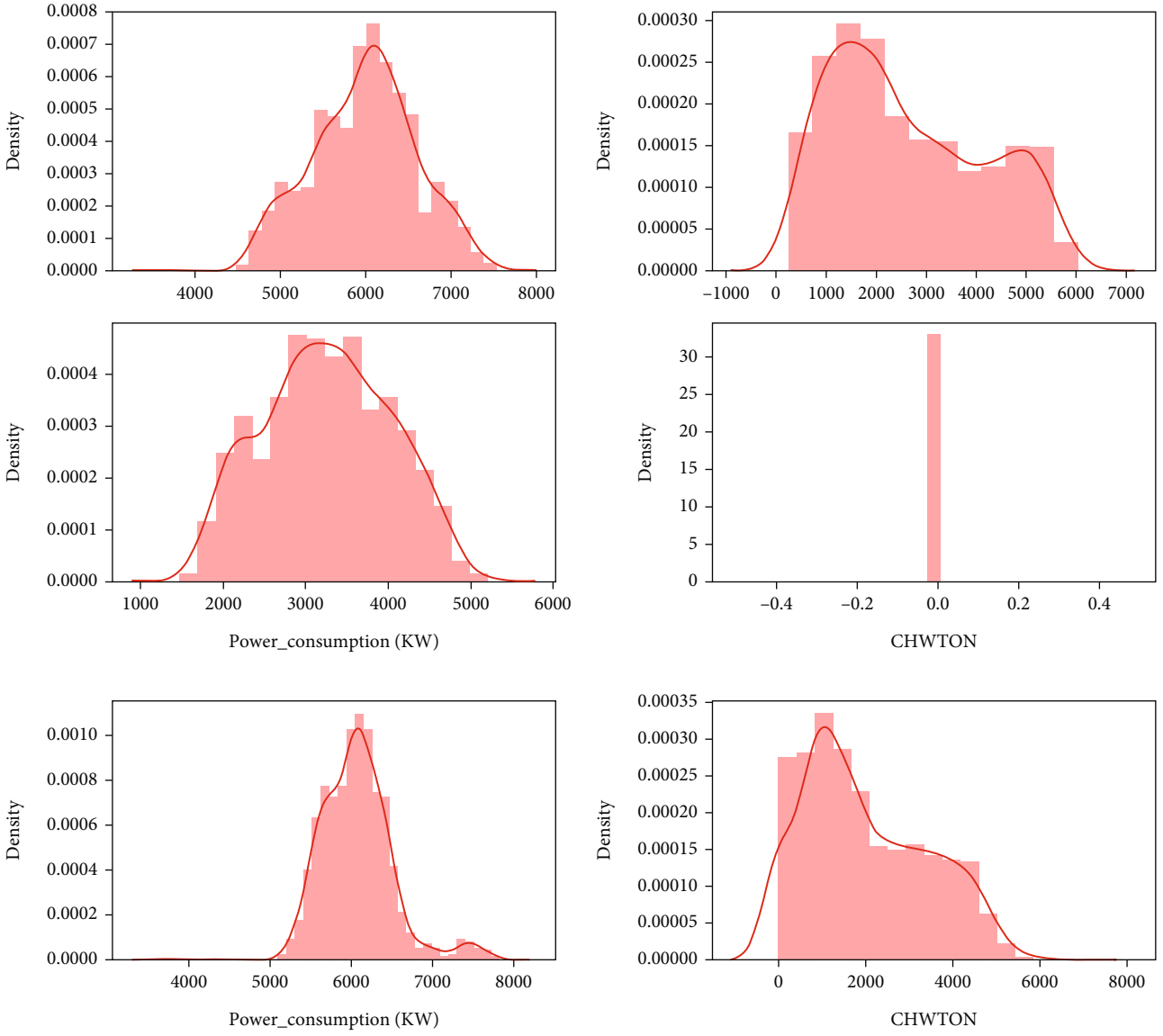

Figure 8: Probability density for buildings CLAS, NHAI, and Cronkite.

platykurtic. This was also evident in Figure 8, where the probability distribution plot had a higher tail and a larger peak center. However, the Cronkite building had a kurtosis value greater than 0, indicating a leptokurtic distribution with higher variance. CLAS and NHAI had roughly normal distributions, but CLAS had a lower mean than the median. Department CLAS also had an almost normal distribution but with higher skewness and kurtosis. The CHWTON data set had a higher variation compared to the other data sets.

<bound method DataFrame.info of campus bldgno

0	Downtown	309	Beus	center	for	law	and	society		2020	1
1	Downtown	309	Beus	center	for	law	and	society		2020	1
2	Downtown	309	Beus	center	for	law	and	society		2020	1
3	Downtown	309	Beus	center	for	law	and	society		2020	1
4	Downtown	309	Beus	center	for	law	and	society		2020	1
...	...	...					...		...	...	...
1092	Downtown	309	Beus	center	for	law	and	society		2022	12
1093	Downtown	309	Beus	center	for	law	and	society		2022	12
1094	Downtown	309	Beus	center	for	law	and	society		2022	12
1095	Downtown	309	Beus	center	for	law	and	society		2022	12
1096	Downtown	309	Beus	center	for	law	and	society		2023	1

	Day	Hour	KW	KWS	...	CHWTONgalsgas	HTmmBTU#Houses	\
0	1		5364.07	−0.01	...	101	542	
1	2		5902.25	−0.01	...	103	517	
2	3		5915.77	−0.01	...	109	529	
3	4		5496.93	−0.01	...	124	511	
4	5		5512.42	−0.01	...	117	479	
...	...	...	...	...	...	...	...	
1092	28		4998.70	−0.01	...	129	451	
1093	29		5012.78	−0.01	...	118	496	
1094	30		4900.29	−0.01	...	150	476	
1095	31		4631.51	−0.01	...	160	507	
1096	1		4694.85	−0.01	...	115	527	

	HTmmBTUlightbulbs	HTmmBTUgalsgas	Total#Houses	Totallightbulbs	\
0	135532	52	3129	782052	
1	129250	49	3283	820649	
2	132119	51	3360	839804	
3	127761	49	3380	844761	
4	119682	46	3282	820307	
...	...	...	...	...	
1092	112759	43	3232	807960	
1093	124047	47	3167	791672	
1094	118887	45	3446	861333	
1095	126654	48	3504	875937	
1096	131700	50	3071	767612	

	Totalgalsgas	GHG	DOW	tstamp2
0	8905	2.550	4	2020-01-01T00:00:00.000
1	9162	2.702	5	2020-01-02T00:00:00.000
2	9651	2.710	6	2020-01-03T00:00:00.000
3	10861	2.452	7	2020-01-04T00:00:00.000
4	10305	2.467	1	2020-01-05T00:00:00.000
...	...	...	...	...
1092	11252	2.669	4	2020-12-28T00:00:00.000
1093	10343	2.567	5	2020-12-29T00:00:00.000
1094	12972	2.558	6	2020-12-30T00:00:00.000
1095	13775	2.346	7	2020-12-31T00:00:00.000
1096	10026	2.508	1	2020-12-32T00:00:00.000

Figure 9: Summary of transform data set for CLAS building.

4.2. Data Preprocessing. Based on Figure 9, the original data set had various scale ranges for power consumption factors like KWS, CHWTON, voltage, and building occupants. To verify the prediction capacity of 29 features, multiple approaches like correlation analysis, ensemble analysis, and tree-based models were used. After testing the mentioned methods, the most ideal qualities for projecting energy demand and consumption are as follows:

(1) Previous consumption patterns

(2) Calendar: weekday, month, and season

(3) Demography: A building's population might influence consumption patterns

(4) Geographical factors such as climate. People will use more electrical appliances at hot and low temperatures, respectively

The study on missing data utilized the missingness matrix to quantify the extent of missing data and identify rows that contained missing values. Upon analyzing Figure 10, it is noteworthy that none of the three data sets exhibited any missing data.

Figure 10: Missingness graph of CLAS building.

4.3. Feature Selection. Selecting the most crucial features plays a vital role in enhancing the effectiveness, stability, and scalability of our prediction model. Through the utilization of a feature importance assessment method, as summarized in Table 4, we identified the top five influential features: KW, KWS, CHWTON, total houses, and CHWTONgaslas. The ranking of these features is illustrated in Figure 11, which shows the order of their importance. Although the initial analysis considered all 29 parameters, the figure only highlights features that significantly contribute to precision, ensuring a streamlined and informative depiction.

The study is aimed at predicting energy consumption in three educational buildings by identifying key parameters. Through feature selection, we have identified key parameters that significantly impact energy usage. These include "CHwton" or chilled water tons which measures the cooling capacity of chilled water systems, representing the heat energy required to melt one ton of ice in 24 hours. Additionally, "KW" denotes the power consumption of electrical equipment and lighting systems within the buildings. "Totallightbulb" denotes the aggregate number of light bulbs or lamps within the buildings, crucial for various assessments. Furthermore, aspects of HVAC systems, like "CHWTONgalsgas," offer insights into chilled water and gas usage. Moreover, "Combined mmBTU" measures the heat required to raise the temperature of water by one degree Fahrenheit. The feature selection process helps identify the most influential parameters for the predictive model, enabling more accurate energy consumption forecasts.

4.4. Performance Evaluation and Comparison. The prediction models' performance was evaluated by comparing multiple methods for each building after training and testing. Comparative results are shown in Table 5.

Table 4: Feature importance.

Features	Importance
CHWTON	0.303456
CHWTONgalsgas	0.23327
Total houses	0.291235
KW	0.353657
HTmmBTU	0.083478
Combined mmBTU	0.183562
HTmmBTUgalsgas	0.285535
Total light bulbs	0.229835

Based on the performance evaluation measurements presented in Table 5, the GBR method exhibited outstanding performance across all buildings. Notably, the determination coefficients were remarkably high, reaching 0.998 for Cronkite, 0.984 for CLAS, and 0.845 for NHAI. Furthermore, the corresponding mean squared error (MSE) values were 8.148, 5.09, and 9.17, respectively. The root mean squared error (RMSE) and mean absolute error (MAE) also supported these results, indicating that GBR outperformed other methods and yielded the best values. Additionally, when assessing the mean absolute percentage error (MAPE) results, GBR surpassed the other methods, demonstrating the lowest error percentage. The LSTM method exhibited lower determination coefficients compared to the GBR results, with values of 0.86 for CLAS, 0.7772 for NHAI, and 0.7609 for Cronkite. However, when comparing LSTM to the RF method, the performance varied across buildings.

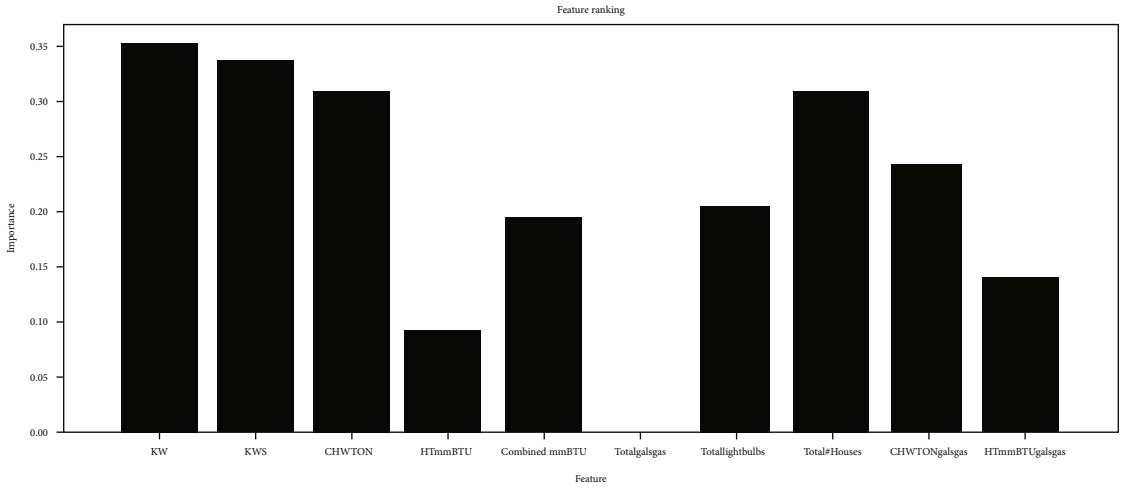

Figure 11: Feature importance.

Table 5: Predictions for performance evaluation using trained models.

Building	Method	R2	MSE	RMS	MAE	MAPE
CLAS	RF	0.8506	27.245	16.530	219.73	76.947
CLAS	LSTM	0.8669	11.0921	13.3036	79.0677	56.0298
CLAS	GBR	0.984	8.148	9.335	71.722	40.2587
NHAI	RF	0.479	56.293	20.439	47.78	8.45123
NHAI	LSTM	0.8372	27.10199	19.2844	33.0788	48.11382
NHAI	GBR	0.795	15.089	17.4370	32.675	52.57254
Cronkite	RF	0.89318	19.821	10.8491	117.704	56.54793
Cronkite	LSTM	0.76096	26.12360	7.3153	29.09945	64.8606
Cronkite	GBR	0.99817	9.1734	4.04234	16.3405	36.34167

Table 6: Real and predicted average consumption for each method.

Real values	GBR test	LSTM test	RF test
5364.07	5511.33	5646.09	5666.75
5902.25	5881.72	5811.137	5822.69
5915.77	5900.722	5871.49	5870.67
5496.93	5491.65	5499.29	5496.55
5512.42	5523.29	5535.02	5535.18
6173.30	6178.63	6366.57	6354.28
6141.73	6296.113	6345.22	6365.09
6302.20	6364.17	6371.89	6389.07
6182.52	6182.480	6234.04	6240.37
6251.62	6171.9	6166.97	6168.348
6251.62	5606.99	5530.53	5527.52
5602,05	5626.398	5637.729	5527.52
5678,72	6769.29	6437.53	5641.79
6842,53	6893.64	6884.76	6417.95
6980,2	6701.16	6606.624	6876.06
6767,83	6695.89	6520.20	6622.32
6568,83	6394.089	6358.6	6506.90
5789,52	5730.96	5596.32934	6355.81

Table 7: Cross-validation score for models.

Algorithm	RF	LSTM	GBR
Validation score	0.83	0.92	0.95

Specifically, in the Cronkite building, the random forest method outperformed LSTM with an R2 value of 0.89. Nevertheless, in terms of other metrics such as MSE and RMSE, LSTM yielded comparably smaller values than the RF method. Moreover, there was a significant difference in the MAPE results, with LSTM generating fewer errors compared to random forest. This observation suggests that, in terms of errors, LSTM performed better and produced a lower number of errors compared to RF. According to the forecast evaluation, the square error method was deemed a more suitable evaluation metric for assessing the accuracy of the predictions. Following this examination, it became clear that the gradient boosting regressor (GBR) method performed the best across all buildings.

Considering the data presented in Table 6, it is evident that the algorithm closest to the real testing values is the gradient boosting regressor, demonstrating good precision. The

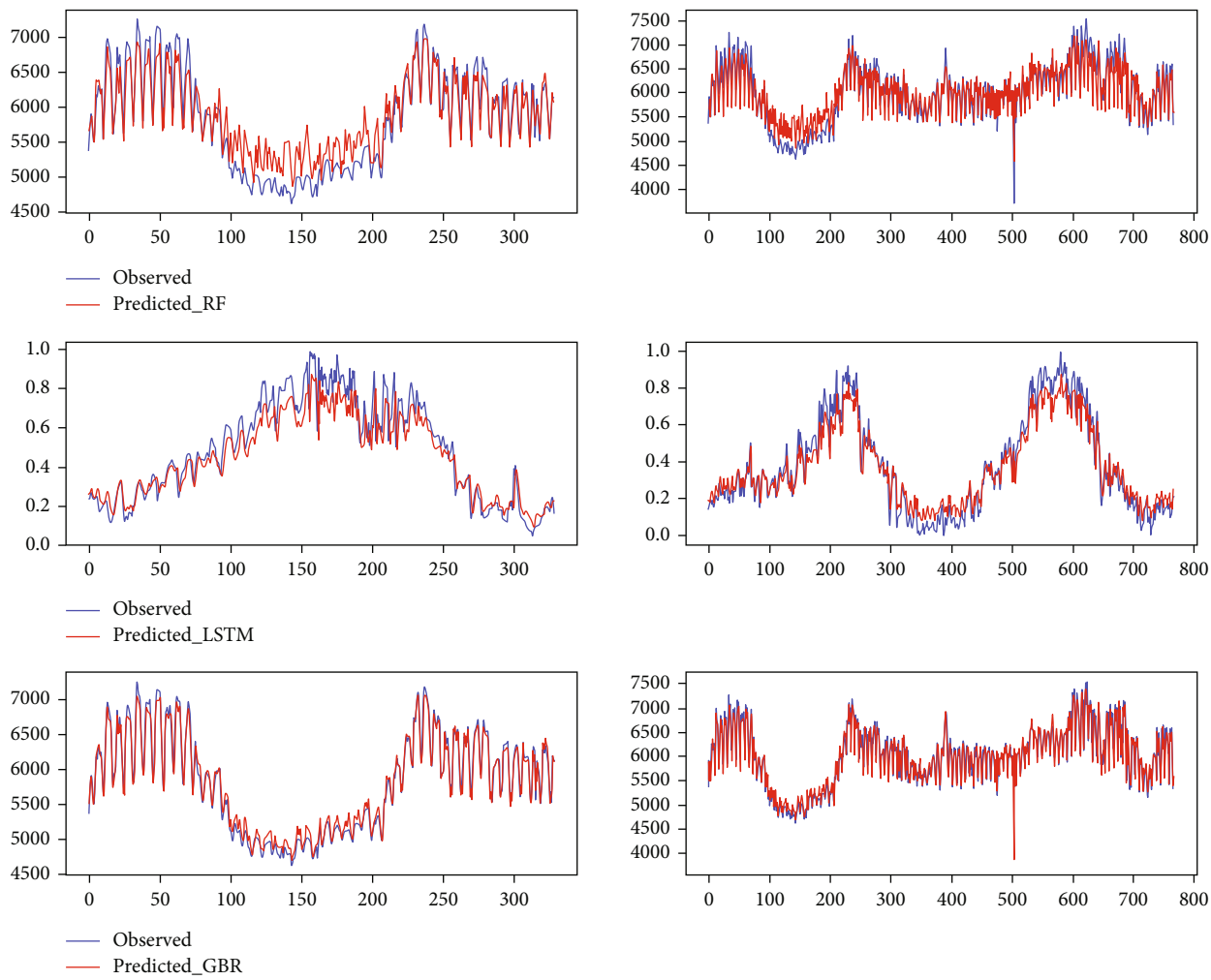

Figure 12: Real and predicted average consumption for CLAS building.

long short-term memory (LSTM) method follows in second place, and the random forest algorithm comes last in terms of accuracy in predicting average consumption. In the context of result validation, K-fold cross-validation is a highly suitable technique for our case due to its inherent advantages. By partitioning the data set into K subsets, each containing a representative sample of the data, K-fold cross-validation ensures thorough training and validation of the model. This approach maximizes data utilization and minimizes bias, as every data point is utilized for both training and validation across different folds. Furthermore, the averaging of performance metrics over multiple splits provides a robust evaluation, effectively reducing the variance associated with a single train-test split. Additionally, K-fold cross-validation facilitates better generalization by assessing the model's performance across diverse subsets of the data, ensuring that it can effectively handle various scenarios. Its utility extends to hyperparameter tuning, enabling the comparison of different parameter configurations across multiple validation sets.

In our scenario, we choose 5-fold cross-validation for its moderate data set size, balancing computational efficiency and robust performance estimation. This method ensures reliable model evaluation without excessive computational overhead and aligns with common practices in the field, allowing easier comparison with existing literature and benchmarks. Table 7 provides the outcome of the 5-fold cross-validation.

A line graph comparison was used to better demonstrate the difference between the actual and anticipated average consumption levels, as depicted in Figures 12–14. In addition, Figures 15–17 show the graphical presentation of the regression line for the three buildings. In the CLAS and Cronkite buildings, the gradient boosting regressor (GBR) produces a symmetric regression line, indicating that its predicted values closely align with the actual ones. Conversely, for the NHAI building data set, characterized by nonsymmetrical data, long short-term memory (LSTM) outperforms other models due to its ability to capture temporal dependencies.

However, in the case of NHAI, the performance difference between LSTM and GBR is minimal, highlighting the suitability of both algorithms for different data characteristics. GBR excels in all cases, while LSTM's recurrent nature makes it valuable for handling nonlinear, time-dependent data.

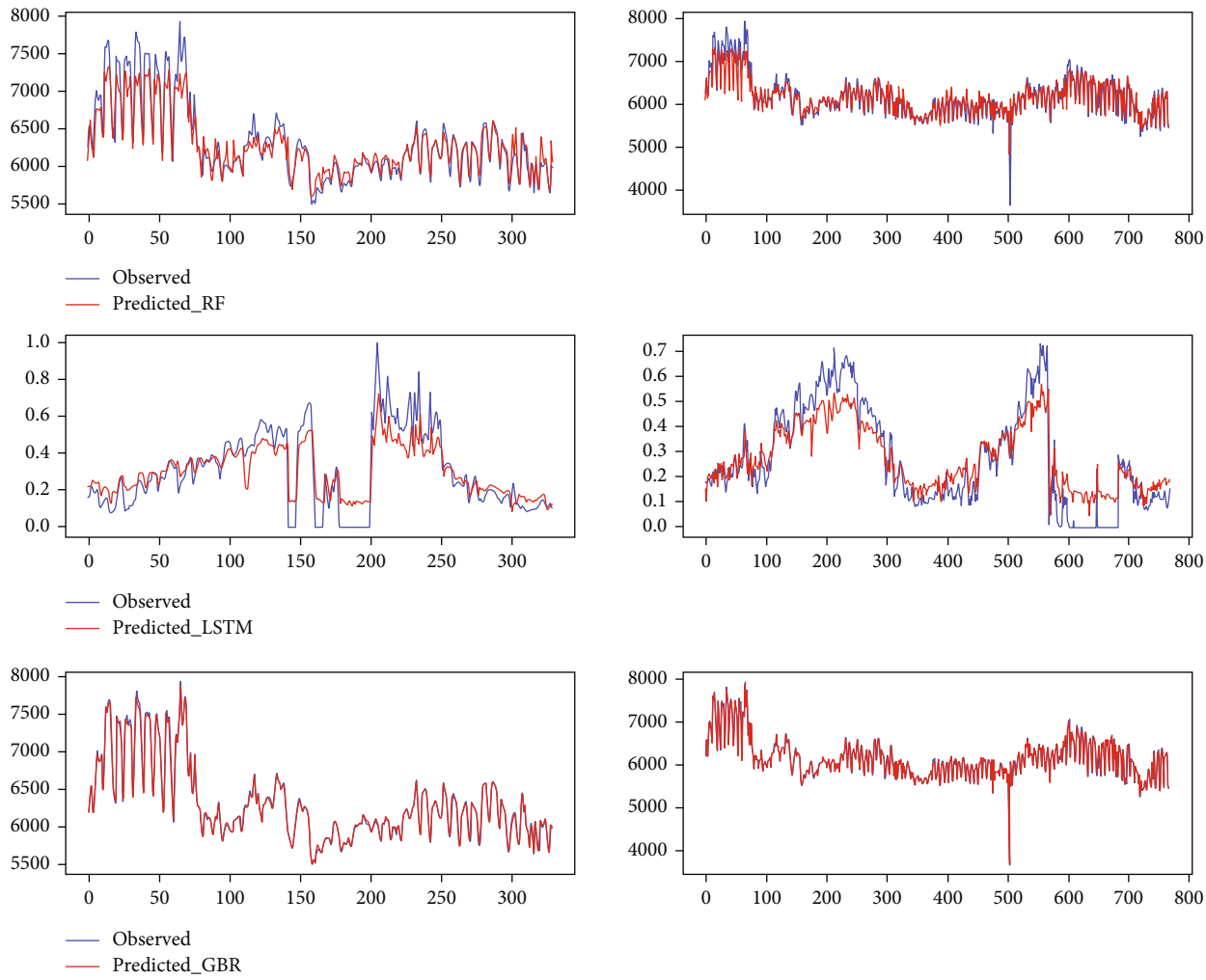

Figure 13: Real and predicted average consumption for Cronkite building.

From the analysis of all the tables and figures, we conclude that the best performances are consistently achieved by the gradient boosting regressor (GBR). GBR's sequential training approach trains weak learners sequentially, correcting errors from previous iterations, and fine-tuning the model's predictive capabilities with each step. Additionally, gradient descent optimization minimizes prediction errors, leading to more accurate predictions. Following GBR, long short-term memory (LSTM) stands out as it is specifically designed for handling sequential data, making it well-suited for time series forecasting and similar tasks. Its ability to understand and process temporal patterns contributes to accurate predictions in time-dependent scenarios. Lastly, the random forest algorithm also delivers good results, particularly when it comes to capturing complex nonlinear correlations between features and the subject variable, and its ability to model complex interactions and patterns makes it effective.

The CLAS building has a significantly higher energy consumption rate, exceeding 30 kWh, in contrast to the other buildings. The main reason for this difference is the large surface area and the simultaneous use for many educational objectives. On the other hand, the Cronkite building has an energy consumption rate of 26 kWh/h, while NHAI has a consumption rate of 12 kWh per hour. Predictive modeling approaches are necessary for efficient energy allocation and management. Within this particular instance, the gradient boosting regressor model demonstrates its superiority in effectively predicting outcomes for both the CLAS and Cronkite buildings. The choice is backed by the model's remarkable performance metrics, as shown by its coefficient of determination (R-squared) values of 0.99 for Cronkite and 0.98 for CLAS. This model improves the accuracy of forecasting by offering proactive insights into the energy needs of each building. It also helps in preventing energy loss before it happens and promotes efforts to reduce energy usage.

5. Comparison with the Previous Study

The study compared three algorithms: random forest, LSTM, and gradient boosting regressor, revealing their performance in forecasting monthly average consumption.

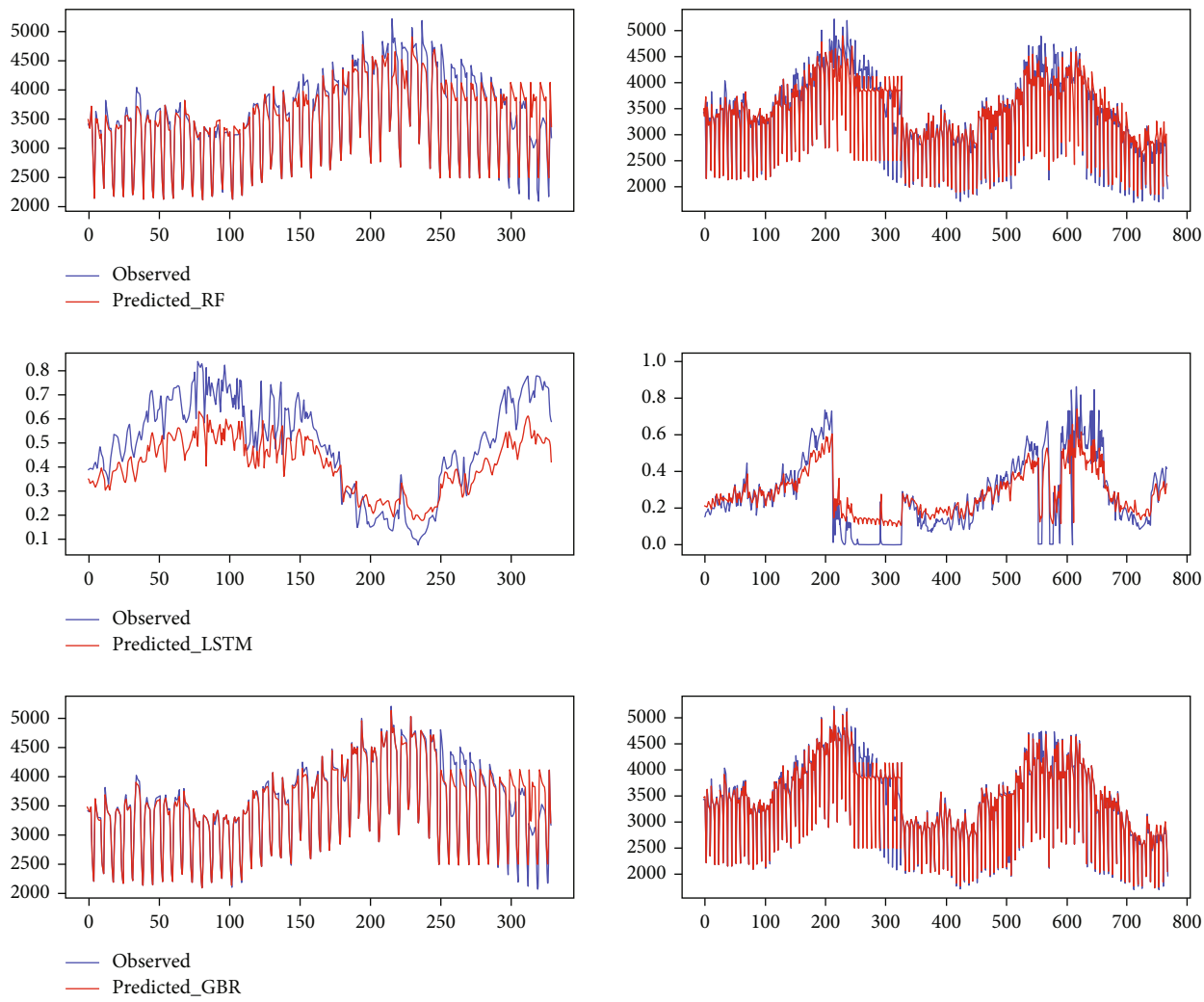

Figure 14: Real and predicted average consumption for NHAI building.

The development of prediction models demonstrated their capabilities, urging further optimization. The findings also led to a comparative analysis with previous machine learning studies. In the first research conducted by Khaoula et al. in 2022 [40], four machine learning algorithms were implemented to predict energy demand for a commercial building over two years. The algorithms used were multiple linear regression (MLR), long short-term memory (LSTM), simple linear regression (LR), and random forest (RF). The results indicated that LSTM performed the best, followed by RF, MLR, and LR, providing valuable insights into the regression algorithms' capabilities. In the second research, Khaoula et al. in 2023 [41] examined energy consumption prediction in a low-energy house over four months. Unlike the first research, this time, the prediction considered not only the house's energy but also its appliances. Three machine learning algorithms, namely, artificial neural networks (ANN), recurrent neural networks (RNN), and random forest (RF), were employed for tests. Recurrent neural networks especially LSTM once again outperformed the other algorithms, achieving an impressive accuracy of 96%. RF was followed with 88% accuracy. However, ANN yielded negative predictions, indicating its unsuitability for time series data sets. Furthermore, in their research, Khaoula et al. [42] used three deep learning algorithms—recurrent neural networks (RNNs), artificial neural networks (ANNs), and autoregressive neural networks (AR-NNs)—to forecast the total load of HVAC systems. The results showed that the autoregressive neural network model outperformed the other two due to its ability to capture temporal dependencies and patterns in time series data, which is crucial for HVAC load prediction. AR-NNs use a simpler architecture, focusing on past observations to predict future values, and their autoregressive nature allows them to effectively model the self-dependence of time series data, leading to more accurate predictions.

Drawing insights from these three studies, significant findings emerge regarding the efficacy of regression algorithms for energy consumption prediction. Specifically, long short-term memory (LSTM) and random forest (RF) consistently emerge as top performers, especially in handling time series data. However, our research introduces a novel aspect by exploring the effectiveness of gradient boosting regressor (GBR), which yielded exceptional results. Notably, GBR

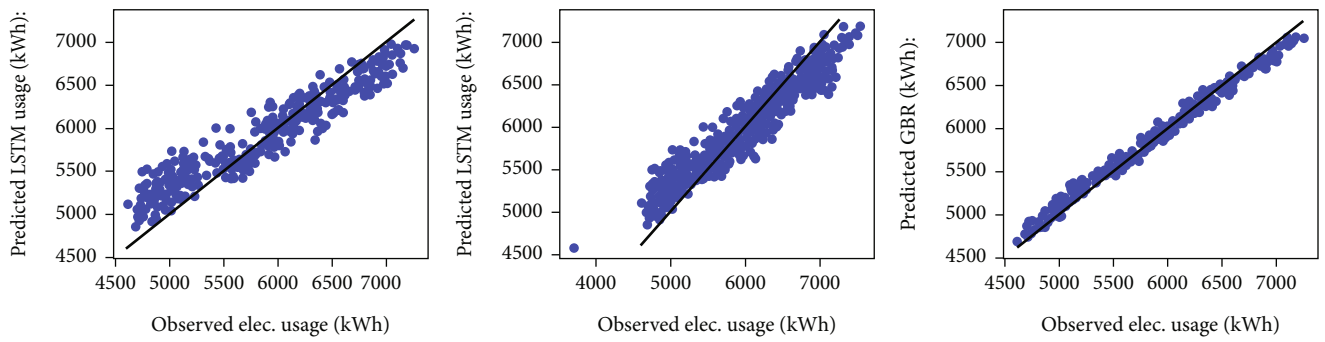

Figure 15: Regression line between observation and predictions for CLAS building.

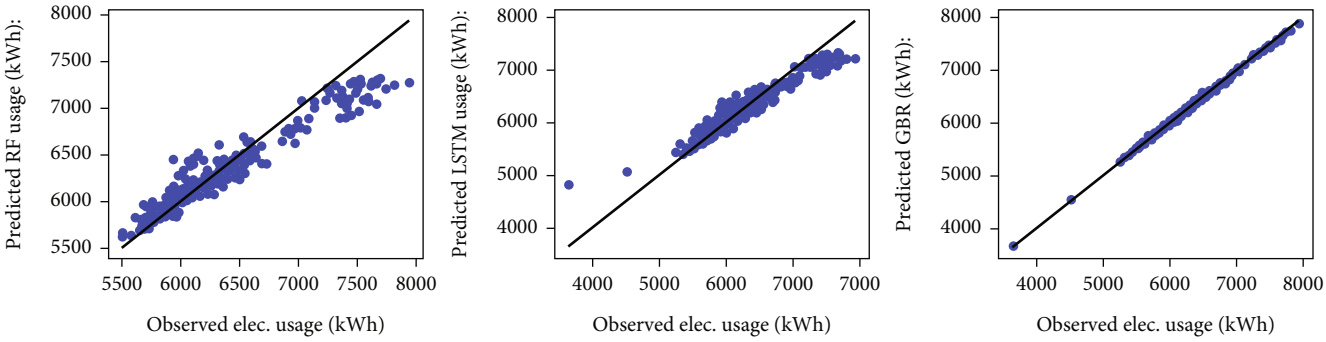

Figure 16: Regression line between observation and predictions for Cronkite building.

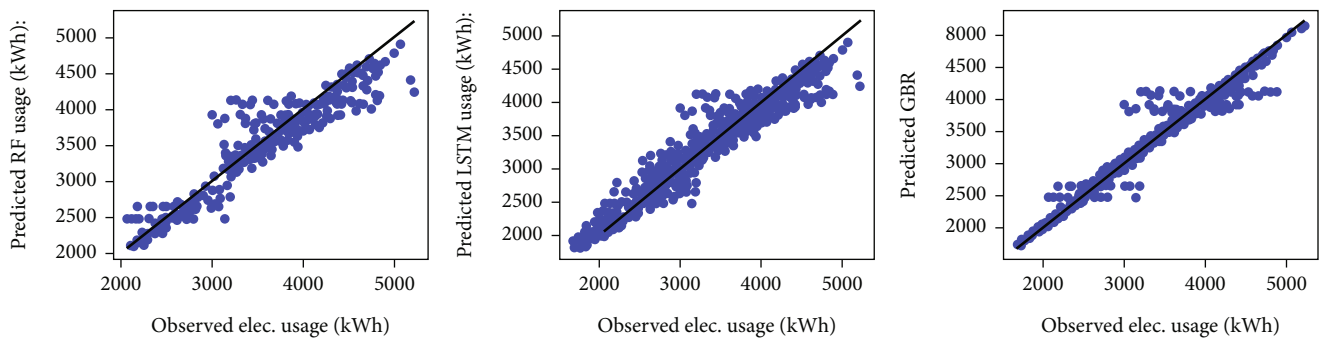

Figure 17: Regression line between observation and predictions for NHAI building.

achieved remarkable precision, boasting an impressive accuracy of 98.9%. Moreover, compared to other algorithms, GBR demonstrated superior performance with fewer errors, as evidenced by lower root mean square (RMS), mean absolute error (MAE), and mean absolute percentage error (MAPE) values. This underscores the potential of GBR as a formidable contender in energy consumption prediction tasks, offering a promising alternative to LSTM and RF in certain contexts.

6. Perspectives and Future Work

For future contributions, we plan to optimize the GBR model by increasing the data used for training and prediction, which may improve efficiency and performance on larger data sets. We intend also to apply a novel approach to the gradient boosting optimizer to fine-tune the model's parameters and hyperparameters more effectively. These efforts are aimed at enhancing the GBR algorithm's performance for accurate energy consumption forecasting and other applications.

Another significant contribution of our future research lies in the utilization of transformer models for predicting diurnal energy consumption patterns. Transformers, originally designed for natural language processing tasks, have shown remarkable capabilities in capturing long-range dependencies in sequential data, making them well-suited for time series forecasting tasks as well. By applying transformer architectures to predict diurnal energy consumption, we aim to leverage their ability to effectively model complex

temporal patterns and dependencies inherent in energy consumption data. Our case study focuses on commercial and institutional buildings, where accurate energy consumption prediction is crucial for optimizing building operations, reducing costs, and minimizing environmental impact.

7. Conclusion

Our major focus in this research is developing an energy consumption forecasting model given the environment of three institutional buildings that have adopted the smart building ecosystem. From January 2020 to January 2023, the collected energy consumption data was subjected to statistical analysis to assess its normality. The skewness and kurtosis values showed that the data had a variety of distribution characteristics.

The predictive model development process involved data preprocessing, which included handling missing data and identifying feature importance. For this research's objective, three supervised machine learning methods, namely, gradient boosting regressor (GBR), long short-term memory (LSTM), and random forest (RF), were selected as the algorithms for the predictive model. The comparison of these strategies was based on an assessment of their production structures and prediction abilities. The results of our model training and testing indicated that each strategy performed differently for each building. Remarkably, the GBR approach continually produced the most promising outcomes, cementing its position as the best-performing strategy across all three buildings: CLAS, NHAI, and Cronkite. GBR's mean absolute percentage error (MAPE) values were 9.337, 12.338, and 4.045 for CLAS, NHAI, and Cronkite, respectively. Additionally, GBR achieved a lower mean absolute error (MAE) for CLAS and Cronkite (71.04 and 53.77, respectively), while RF and LSTM yielded lower MAE results for these two buildings. Moreover, while computing average consumption using demand data, it was shown that the gradient boosting regressor (GBR) displayed greater accuracy in anticipating demand. This performance outperformed all other approaches in all buildings.

In terms of future study recommendations, it is suggested to use more powerful computers or platforms to run the LSTM algorithm, potentially improving its performance. Additionally, exploring hybrid or ensemble methods may be beneficial, as they have shown higher accuracy than single regressors. Lastly, a comparison with another smart building could be included to distinguish and validate the obtained results. These recommendations can further enhance the understanding and applicability of the energy consumption predictive model.

Data Availability

The collected data was saved in an open-source website server [43] and could be manually downloaded from the platform's website in the form of a CSV file with any sort of aggregation [43] (https://portal.emcs.cornell.edu/d/2/dashboard-list?orgId=2).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the National Center for Scientific and Technical Research (CNRST) for supporting and funding this research.

References

[1] H. Farzaneh, L. Malehmirchegini, A. Bejan, T. Afolabi, A. Mulumba, and P. P. Daka, "Artificial intelligence evolution in smart buildings for energy efficiency," *Applied Sciences*, vol. 11, no. 2, p. 763, 2021.

[2] E. Khaoula, B. Amine, and B. Mostafa, "Machine Learning and the Internet Of Things for Smart Buildings: A state of the art survey," in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1–10, Meknes, Morocco, 2022.

[3] A. Widya-Hasuti, A. Mardani, D. Streimikiene, A. Sharifara, and F. Cavallaro, "The role of process innovation between firm-specific capabilities and sustainable innovation in SMEs: empirical evidence from Indonesia," *Sustainability*, vol. 10, no. 7, p. 2244, 2018.

[4] MarketsandMarkets, *Smart Building Market by Component (Solution and Services), Solution (Safety and Security Management, Energy Management, Building Infrastructure Management, Network Management, and IWMS), Services, Building Type*, Region - Global Forecast to 2025, 2021.

[5] K. B. Anacker, *Healthy Buildings: How Indoor Spaces Drive Performance and Productivity*, J. G. Allen and J. D. Macomber, Eds., Harvard University Press, Cambridge, 2020.

[6] American Council for an Energy-Efficient Economy, "Building technologies," n.d., https://www.aceee.org/topics/building-technologies.

[7] US Department of Energy, "Energy efficient commercial buildings," n.d., https://www.energy.gov/eere/buildings/energy-efficientcommercial-buildings.

[8] Intel, "Intelligent buildings: saving energy, making occupants happy," 2017, https://www.intel.com/content/www/us/.

[9] Z. Chen, C. Lin, X. Zhou, L. Huang, M. Sandanayake, and P.-S. Yap, "Recent technological advancements in BIM and LCA integration for sustainable construction: a review," *Sustainability*, vol. 16, no. 3, p. 1340, 2024.

[10] P. IEA, *World Energy Outlook 2022*, International Energy Agency (IEA), Paris, France, 2022.

[11] "Smart building market- by application (commercial, residential, and industrial), by automation type (energy management, infrastructure management, network and communication management), by service (professional service and managed services) and by region global industry perspective, comprehensive analysis, and forecast 2021-2028," https://www.zionmarketresearch.com/report/smart-buildingmarket.

[12] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper, "Machine learning for estimation of building energy consumption and performance: a review," *Visualization in Engineering*, vol. 6, pp. 1–20, 2018.

[13] M. A. Ahajjam, D. B. Licea, C. Essayeh, M. Ghogho, and A. Kobbane, "MORED: a Moroccan buildings' electricity consumption dataset," *Energies*, vol. 13, no. 24, p. 6737, 2020.

[14] N. Somu, M. R. Gauthama Raman, and K. Ramamritham, "A hybrid model for building energy consumption forecasting using long short term memory networks," *Applied Energy*, vol. 261, article 114131, 2020.

[15] I. W. A. Suranata, I. N. K. Wardana, N. Jawas, and I. K. A. A. Aryanto, "Feature engineering and long short-term memory for energy use of appliances prediction," *TELKOMNIKA (Telecommunication Computing Electron- ics and Control)*, vol. 19, no. 3, pp. 920–930, 2021.

[16] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: case study in Malaysia," *Developments in the Built Environment*, vol. 5, article 100037, 2021.

[17] M. Faiq, K. G. Tan, C. P. Liew et al., "Prediction of energy consumption in campus buildings using long short-term memory," *Alexandria Engineering Journal*, vol. 67, pp. 65–76, 2023.

[18] T. Kawahara, K. Sato, and Y. Sato, "Battery voltage prediction technology using machine learning model with high extrapolation accuracy," *International Journal of Energy Research*, vol. 2023, Article ID 5513446, 17 pages, 2023.

[19] S. A. Mohammed, O. A. Awad, and A. M. Radhi, "Optimization of energy consumption and thermal comfort for intelligent building management system using genetic algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 3, pp. 1613–1625, 2020.

[20] N. F. Aurna, M. T. M. Rubel, T. A. Siddiqui et al., "Time series analysis of electric energy consumption using autoregressive integrated moving average model and Holt Winters model," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, pp. 991–1000, 2021.

[21] Z. Ferdoush, B. N. Mahmud, A. Chakrabarty, and J. Uddin, "A short-term hybrid forecasting model for time series electrical-load data using random forest and bidirectional long short-term memory," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 763–771, 2021.

[22] Y. He and K. F. Tsang, "Universities power energy management: a novel hybrid model based on iCEEMDAN and Bayesian optimized LSTM," *Energy Reports*, vol. 7, pp. 6473–6488, 2021.

[23] X.-B. Jin, W.-Z. Zheng, J.-L. Kong et al., "Deep-learning forecasting method for electric power load via attention-based encoder decoder with Bayesian optimization," *Energies*, vol. 14, no. 6, p. 1596, 2021.

[24] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, and S. Ajayi, "Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques," *Journal of Building Engineering*, vol. 45, article 103406, 2022.

[25] J. Jang, J. Han, and S.-B. Leigh, "Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks," *Energy and Buildings*, vol. 255, article 111647, 2022.

[26] A. N. Ndife, W. Rakwichian, P. Muneesawang, and Y. Mensin, "Smart power consumption forecast model with optimized weighted average ensemble," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, p. 1004, 2022.

[27] V. H. Duong and N. H. Nguyen, "Machine learning algorithms for electrical appliances monitoring system using open-source systems," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, p. 300, 2022.

[28] C. Vennila, T. Anita, T. Sri Sudha et al., "Forecasting solar energy production using machine learning," *International Journal of Photoenergy*, vol. 2022, Article ID 7797488, 7 pages, 2022.

[29] S. Kapp, J.-K. Choi, and T. Hong, "Predicting industrial building energy consumption with statistical and machine learning models informed by physical system parameters," *Renewable and Sustainable Energy Reviews*, vol. 172, article 113045, 2023.

[30] R. Bhol, S. C. Swain, R. Dash, K. J. Reddy, C. Dhanamjayulu, and B. Khan, "Short-term reactive power forecasting based on real power demand using Holt-Winters' model ensemble by global flower pollination algorithm for microgrid," *International Journal of Energy Research*, vol. 2023, Article ID 9733723, 22 pages, 2023.

[31] M. M. Asiri, G. Aldehim, F. A. Alotaibi, M. M. Alnfiai, M. Assiri, and A. Mahmud, "Short-Term Load Forecasting in Smart Grids Using Hybrid Deep Learning," *IEEE Access*, vol. 12, pp. 23504–23513, 2024.

[32] "Real-time electricity consumption," https://portal.emcs.cornell.edu/d/2/dashboard-list?orgId=2.

[33] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Annals of Cardiac Anaesthesia*, vol. 22, no. 1, pp. 67–72, 2019.

[34] M. R. Segal, *Machine learning benchmarks and random forest regression*, 2004.

[35] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[36] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.

[37] A. Di Bucchianico, "Coefficient of determination (R^2)," *Encyclopedia of Statistics in Quality and Reliability*, 2008.

[38] D. Wallach and B. Goffinet, "Mean squared error of prediction as a criterion for evaluating and comparing system models," *Ecological Modelling*, vol. 44, no. 3-4, pp. 299–306, 1989.

[39] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," *International Journal of Forecasting*, vol. 15, no. 4, pp. 405–408, 1999.

[40] E. Khaoula, B. Amine, and B. Mostafa, "Evaluation and comparison of energy consumption prediction models," in *International Conference on Advanced Technologies for Humanity*, Marrakech, Morocco, 2022.

[41] E. Khaoula, B. Amine, and B. Mostafa, "Evaluation and comparison of energy consumption prediction models case study: smart home," in *The International Conference on Artificial Intelligence and Computer Vision*, pp. 179–187, Springer Nature Switzerland, Cham, 2023.

[42] E. Khaoula, B. Amine, and B. Mostafa, "Forecasting diurnal Heating Energy Consumption of HVAC system using ANN, RNN, ARNN," in *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pp. 1–6, Casablanca, Morocco, 2023.

[43] "Real time utility use data," https://portal.emcs.cornell.edu/d/2/dashboard-list?orgId=2.