Advanced techniques for electricity consumption prediction in buildings using comparative correlation analysis, data normalization, and Long Short-Term Memory (LSTM) networks: A case study of a U.S. commercial building

Energy Reports 14 (2025) 56–65

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Advanced techniques for electricity consumption prediction in buildings using comparative correlation analysis, data normalization, and Long Short-Term Memory (LSTM) networks: A case study of a U.S. commercial building

Moon Keun Kim [a,*], Yang-Seon Kim [b], Nuodi Fu [c], Jiying Liu [d], Junqi Wang [c], Sanghyuk Lee [e], Jelena Srebric [f]

[a] Department of Built Environment, Oslo Metropolitan University, Oslo N-0130, Norway
[b] Mechanical Engineering Department, College of Engineering, Wichita State University, Wichita, KS 67260, USA
[c] School of Architecture, Southeast University, 2 Sipailou, Nanjing 210096, China
[d] School of Thermal Engineering, Shandong Jianzhu University, Jinan 250101, China
[e] School of Computing, New Uzbekistan University, Tashkent 100007, Uzbekistan
[f] Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Building energy forecast, Data normalization, Deep learning
Building electricity consumption
Weather data

ABSTRACT

This study introduces innovative assessment techniques to comprehend the effect of six data normalization methods, implemented through the LSTM algorithm, on predicting electricity consumption in commercial buildings. The focus lies on analyzing the relationship between various normalization process and its integration with LSTM method concerning building electricity consumption. The LSTM model incorporates input nodes from diverse sources, including weather data, plug load data, and occupancy ratio. Six distinct normalization process—Min-Max, Mean, Z-score, Gaussian, VSS, and IQR—are applied to assess the model's evaluation on both training and test datasets. The study found that combining the LSTM method with Min-Max and IQR achieves lower figures, representing better performance and greater stability in comparison to alternative normalization techniques. These results described the critical role of data normalization in improving the performance of LSTM models, highlighting the importance of choosing suitable normalization techniques for specific applications while balancing improved accuracy against computational complexity. Furthermore, the study explores the correlation impact of variations in average input elements, particularly with a 5% increase and decrease in value, on electricity consumption in commercial buildings across different seasons. Plug loads emerge as dominant contributors to electricity consumption across various normalization methods and raw data, with temperature and humidity ratio exerting notable influence in specific seasons.

1. Introduction

Anticipating the demand for electricity is of utmost importance in various processes (da Silva and Meneses, 2023). This predictive methodology is an essential tool for making decisions about the energy load in smart grid systems (Caro et al., 2020). It guarantees stability in electricity generation systems to achieve maximum efficiency and effectively manages energy demands in the ever-changing energy market (Bunn, 2000). Furthermore, a very precise forecasting model is essential for identifying anomalies (Himeur et al., 2021) especially when evaluating the ratios of building occupancy and the surrounding environmental conditions. This concept is crucial in reducing energy losses and waste (Cheng et al., 2022).

The growing prevalence of data-driven analyses in building energy forecast approaches is shown in multiple study disciplines. These data-driven methods, which often need fewer input elements, are relatively simple yet exhibit impressive accuracy. Occasionally, they surpass conventional engineering-based energy forecasting approaches. (da Silva and Meneses, 2023; Moon Keun Kim et al., 2020; Kim et al., 2024; Li et al., 2022; Runge and Zmeureanu, 2019).

* Corresponding author.
E-mail address: Moon.Kim@oslomet.no (M.K. Kim).

https://doi.org/10.1016/j.egyr.2025.05.074
Received 18 February 2025; Received in revised form 11 May 2025; Accepted 26 May 2025
Available online 6 June 2025

Building electricity consumption sectors have extensively employed various data-driven algorithms. Machine learning models, with their performance to deal with non-linear patterns, offer flexibility that is often lacking in conventional statistics methods (Zhang et al., 2024). Artificial Neural Network (ANN) methods have been extensively performed in several areas, for implementing projecting usage demand, evaluating the present electricity usage pattern, and identifying strategies for smart building management techniques (C. J. Lu et al., 2022; Peplinski et al., 2024; Wang et al., 2023).

Temporal dependencies within time series data are effectively addressed by the Recurrent Neural Networks (RNN) methodology, utilizing recurrent networks to recalculate values from previous time steps for estimating building energy consumption. Numerous studies have substantiated the efficacy of RNN as a valuable tool for building energy forecasting based on time series (da Silva and Meneses, 2023; C. Lu et al., 2022).

The Long Short-Term Memory Networks (LSTM) method stands as an advanced version of the deep RNN (Hasan, 2023; Peng et al., 2022; Wu et al., 2025). Traditional RNNs face challenges in training with back-propagation due to rapid reduction of errors, leading to vanishing gradients that impede or halt the learning process (Schmidhuber, 2015). In contrast, the LSTM algorithm was specifically formed to address the vanishing gradient issue. Its structure is similar to conventional RNNs, but it replaces summation units in the hidden layer with memory blocks, forming recurrently connected subnets (Graves, 2012; Jinmiao and Chaudhari, 2005). Numerous studies in the literature highlight the adaptability and accurate performance of the LSTM model in energy prediction scenarios.

Data normalization is a preprocessing technique that rescales or alters data in order to ensure that each element contributes equally to the predictive analysis (Kim et al., 2024; D. Singh & B. Singh, 2020). Data normalization is a process that rescales values to a standardized scale, ensuring that no one feature has a disproportionate influence on the analysis due to its magnitude (Ren and Cao, 2020). This technique improves the convergence and ability of ANN process by equalizing the impact of each feature, thereby effectively contributing to the entire analysis (El Bouchefry and de Souza, 2020). Data normalization is essential for creating precise predictive models, and its importance has been widely studied in across Artificial Intelligence (AI) methods.

This study presents an improved analysis that examines the relationship between variables such as occupancy rate, plug-load data, and surrounding environmental conditions with building power usage during different seasons. We evaluated the substantial effect of various data normalization techniques on energy forecasting model in a commercial office building. The suggested predictive control technique relies on ANNs more specifically the LSTM. The study utilizes trained ANNs to estimate the real power usage of a commercial building. The ANNs are trained using specific impact factors such as occupancy rate, plug load data, and surrounding climate factors (solar irradiation, wind velocity, ambient outdoor dry bulb temperature, and specific humidity ratio). The prediction is made using the LSTM algorithm. The primary contents and aims of our investigation include:

- Enhanced analyze the variations using six distinct data normalization algorithms using measurement data achieved in a commercial building, combining occupancy rate, surrounding environment, and electricity usage.
- Forecasting electrical energy usages in a commercial structure utilizing the ANN methods: LSTM for comparative assessment.
- Examine the impact of six standardization methods—Min-Max (minimum and maximum), Variable Stability Scaling (VSS), IQR, Mean, Z-score, and Gaussian normalization techniques—by comparing them to raw data. Apply forecasting analysis utilizing ANN algorithms utilizing normalized data to evaluate its effectiveness in forecasting electricity usages.
- Enhanced correlation analysis techniques with and without data normalization methods within seasonal changes

The primary contribution of this study is offering meaningful insights to enhance the performance of the LSTM-based forecasting model using six distinct data normalization algorithms for electricity usage forecasting in a commercial building in U.S., thereby supporting variable implementation. Initially, this article outlined the following contents: ANNs models and data normalization methods, correlation methods, results, and conclusions.

2. ANNs methods to forecast electricity usages

ANN models have gained popularity in the field of energy usage and system modeling as a favored option for data-driven correlation examination and forecasting methodologies (Soori et al., 2023; Zhu et al., 2022). ANNs excel in effectively managing intricate and non-linear connections among actual data measured, including factors like building occupant ratio, plug load data, dry bulb temperature, specific humidity ratio, solar irradiation, and wind velocity. This leads to greater precision in comparison to conventional statistical or simulation analytical techniques (Barone et al., 2021; Kim et al., 2024; Kim and Srebric, 2017). The benefit of ANNs stems from its architecture, which is derived from the structure and functioning of the human brain (Wang and Srinivasan, 2017). The chosen algorithm for the investigation is the LSTM. The LSTM algorithm is very beneficial for predicting energy demand in variable building types because of its ability to preserve important training data from the start of the series, thereby overcoming the prevalent problem of gradient vanishing observed in RNNs (Jinmiao and Chaudhari, 2005; Weerakody et al., 2021). Numerous studies have reported the LSTM's high accuracy performance in the literature (Zhang et al., 2024). Incorporating the LSTM algorithm into the research improves forecasting and execution capabilities, creating it a key and suitable method for reaching the study aims.

2.1. LSTM neural networks

In contrast to feed-forward neural network (FFNN) algorithm, Recurrent neural networks (RNNs) create links that enable neurons to convey information among layers through recurrent loops in training procedure. This characteristic allows RNNs to excel at identifying and simplifying data types. The feedback mechanism in RNNs enhance accuracy of inputs and output by using calculated outputs to improve the input elements as referenced (Kim and Cho, 2019; Kumar Dubey et al., 2021; Li et al., 2025). LSTM algorithm, a specific type of RNN, distinguish themselves with a unique memory cell capability that enhances information retention over prolonged periods (Hochreiter and Schmidhuber, 1997). In order to overcome the difficulties associated with vanishing and exploding gradients, LSTM algorithm utilize advanced mechanisms including input and forget gates (C. J. Lu et al., 2022; Masood and Ahmad, 2021; Wang et al., 2025). The gates provide improved regulation of gradient flow, leading to greater retention of long-term dependence (Chandra and Matuska, 2022).

These following equations illustrate the LSTM algorithm (Kumar Dubey et al., 2021; Li et al., 2025):

$$\text{Input gate}: i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{1}$$

$$\text{Forget gate}: f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2}$$

$$\text{Cell state updddate}\hat{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \quad \times C_{t-1} + i_t \times \hat{C}_t) \tag{4}$$

$$\text{Output gate}: o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$\text{Hidden State}: h_t = o_t \times \tanh(C_t) \quad (6)$$

The LSTM algorithm includes weight matrices (Hou et al., 2025), W_i, W_f, W_o and W_c to regulate the input, forget, output gate, as well as cell state candidate, consequently, the bias vectors; b_i, b_f, b_o and b_c are connected to these control gates. These activation functions employed in the LSTM element is categorized into the sigmoid function, $\sigma(*)$ and hyperbolic tangent, tanh(*).

Fig. 1 provides a visual concept of the LSTM process algorithms. It depicts the design and element of the LSTM model, emphasizing all gates and cell state. The figure also illustrates the interconnections and information flow among these elements.

3. Data normalization methods

The increasing volume and elements of data pose a significant challenge in machine learning (ML) performance, leading to issues such as data instability due to errors, absent data entries or incomplete values, fluctuates, and data discrepancies (Rousseeuw and Hubert, 2011; Saha and Srivastava, 2014; Singh and Singh, 2022). Literature has addressed the problems of prominent features, which negatively impact the effectiveness in classifying data demonstrated by different learning methods. The presence of these issues arises when data is collected from different databases or combined sources. Outliers affect the accurate estimation of statistics, leading to under-estimation or over-estimation, while main features generate bias in learning, causing classifiers both underfitting and overfitting tendencies on the training data. (Dalwinder Singh & Birmohan Singh, 2020).

Data normalization is a preprocessing method utilized to standardize or modify data, ensuring that each characteristic contributes evenly to the evaluation of performance (Liu et al., 2019). This process entails standardizing the values of various of multiple features to a standardized scale, hence mitigating the potential bias caused by a single variable with a significantly greater extent (Swift et al., 2023). Normalization can be essential for enhancing the efficiency and convergence of ANN methodologies. It ensures that each feature has a balanced influence on the total analysis and predicted results (Singla et al., 2022). Data normalization is vital in the creation of accurate predictive models for energy system and has been thoroughly studied in the relation of several ANN learning algorithms. It has a crucial function in preprocessing input data for modeling by reducing the influence of varying sizes and distributions of characteristics (Zhang et al., 2014).

This research provided six alternative normalizing strategies to optimize the use of ANNs LSTM models and enhanced the predicted accuracy of building energy usage across six measured components. The purpose of selected normalization method is to reorganize the input data, with the goal of enhancing the overall effectiveness and accuracy of the ANN algorithms. Every normalization technique is tailored to particular data attributes, customizing the input elements to better align with the ANN models.

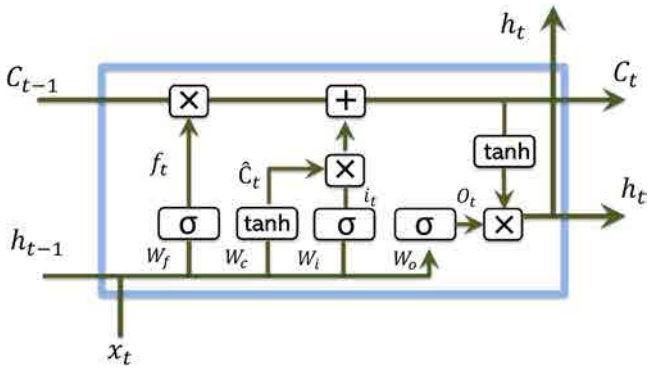

Fig. 1. Visual concept of the LSTM model.

3.1. Min-max (minimum and maximum) data normalization method

Min-Max data normalization utilizes the approach linearly transforms the un-normalized data to fit within pre-established lower and upper bounds rescaled within (e.g., [0,1] or [-1, 1]).

$$x' = \frac{x - \min(\max_n - \min_n)}{\max - \min} + \min_n \quad (7)$$

Where x is the value of original, x' is the value after normalization, min and max represents the minimum and maximum features of the original dataset. The designated lower and upper limits for rescaling the data are indicated by $\max_n$ and $\min_n$ which is normalized dataset.

Min-max data standardization is a method employed for standardizing the influence of variables, ensuring they carry equal weight in the decision-making procedure in Artificial intelligent modeling and data driven examination (Singla et al., 2022).

3.2. Mean data normalization method

Mean standardization is a technique used to scale features in a dataset. The technique entails computing and deducting the average value for each characteristic. Moreover, it is customary to divide this computed value by either the range or the standard deviation (Singla et al., 2022).

The formula of Mean normalization for an element x' is as shown:

$$x' = \frac{x - \text{mean}(x)}{\max - \min} \quad (8)$$

Where x is the original value, x' is the value after normalization, min and max are the minimum and maximum values of the data. The designated lower and upper limits for rescaling the data are indicated by max and min and mean (x) is the mean value of data.

3.3. Z-score data normalization method

This strategy, typically referred to as standardization, is an extensively utilized data normalizing approach in both artificial intelligent and statistical research. This approach involves converting the dataset in order to obtain a mean value of zero and a standard deviation of one for each specific attribute (Liu et al., 2019). Z-score normalization entails transforming each value in a dataset such that the mean of all values is shifted to 0, and the standard deviation is adjusted to 1.

The equation for a feature x' is as illustrated:

$$x' = \frac{x - \text{mean}(x)}{\text{Std(x)}} \quad (9)$$

where x is the value of original, x' denote the standardization value, mean(x) signify the mean of the feature x, and std(x) indicate the standard deviation of x.

3.4. Gaussian function data normalization method

This model depicts the probability features of element from a properly distributed parameters randomly (Liu et al., 2019):

$$x' = \frac{1}{\text{std(x)}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x - mean(x))^2}{std(x)^2}\right) \quad (10)$$

where x represents the value of original, x' denote the standardized value, mean(x) signify the mean of the feature x, and std(x) indicate the standard deviation of x.

The Gaussian normalization finds extensive application in diverse fields such as statistics, theoretical probability, and data-driven evaluation. It is employed to design continuous random features demonstrating characteristics of general distribution (Singla et al., 2022).

3.5. Variable Stability Scaling (VSS)

This VSS strategy broadens the utilization of the Z-score method by include the Coefficient of Variation (CV) as a scaling factor (Singh and Singh, 2022). The Coefficient of Variation is computed by dividing the mean of the data by its standard deviation, as specified below (Singh and Singh, 2022):

$$x' = \frac{(x - \text{mean}(x))}{\text{Std}(x)} \cdot \frac{\text{mean}(x)}{\text{Std}(x)} \tag{11}$$

The CV represents more weight to features with a low standard deviation and reduces the value of features with a high standard deviation.

3.6. Relative interquartile normalization (IQR)

Quartile normalization entails arranging the data in ascending order, splitting it into four equal halves called quartiles, and then determining the median of each quartile. The acquired medians are utilized to normalize the data. The relative interquartile normalization is one of quartile normalization methods and the formular is shown below:

$$x' = \frac{x}{\text{IQR}} \tag{12}$$

where IQR is the value = Q_3- Q_1, Q1 is the lower quartile which corresponds with 25th percentile, Q2 is the median value, Q3 is the upper quartile which is corresponds with 75th percentile.

Fig. 2 illustrates the training and prediction stages of the LSTM network, each employing a different normalization technique utilized in this research. The image provides a clear and logical depiction of the sequential operations entailed in the training and simulating phases for each neural network strategy, along with its corresponding standardization methods. These procedures include data pretreatment, supplied data submission, weight optimization as well as output forecasting.

3.7. Correlation method

To assess the enhanced correlation method among the input parameters measured or collected, this research employed a correlated impact factor (CIF). Disparities in these values signify the extent of effeteness, where positive or negative amount indicate the directionality of the results' effects. (M. K. Kim et al., 2020). As per certain studies, the impact of modifying each node element is determined by calculating correlation impact factor through statistical methods following the deep learning process of ANNs (M. K. Kim et al., 2020; Kim et al., 2024; Ye and Kim, 2018).

The procedure for computing the CIF is as shown: Upon completing the training procedure of ANNs, the testing values undergo a 5 % increase or decrease from their initial values to generate artificial testing data. This 5 % variation is considered as the influence of each parameter, reflecting a linearity association among the altered outcomes. Consequently, the outcomes generated by the modified testing nodes are compared to the actual results, utilizing the initial testing data value as a reference for forecasting.

The CIF value is determined by assessing the variation in magnitude between the predicted data. These results serve as a means to evaluate the effectiveness of each element on the overall electricity usages (Kim et al., 2019).

$$\textit{Correlated Impact factor}(CIF) = \frac{\frac{y_{\textit{test results with increasing or decreasing5\%of sample}} - y_{\textit{test results}}}{y_{\textit{test result}}}}{0.05} \tag{13}$$

To assess the precision and error rate of the LSTM algorithm employing six normalization techniques, this research utilized the coefficient of variation of the normalized mean bias error (NMBE) as well as the root mean square error (CVRMSE). The equations are followed

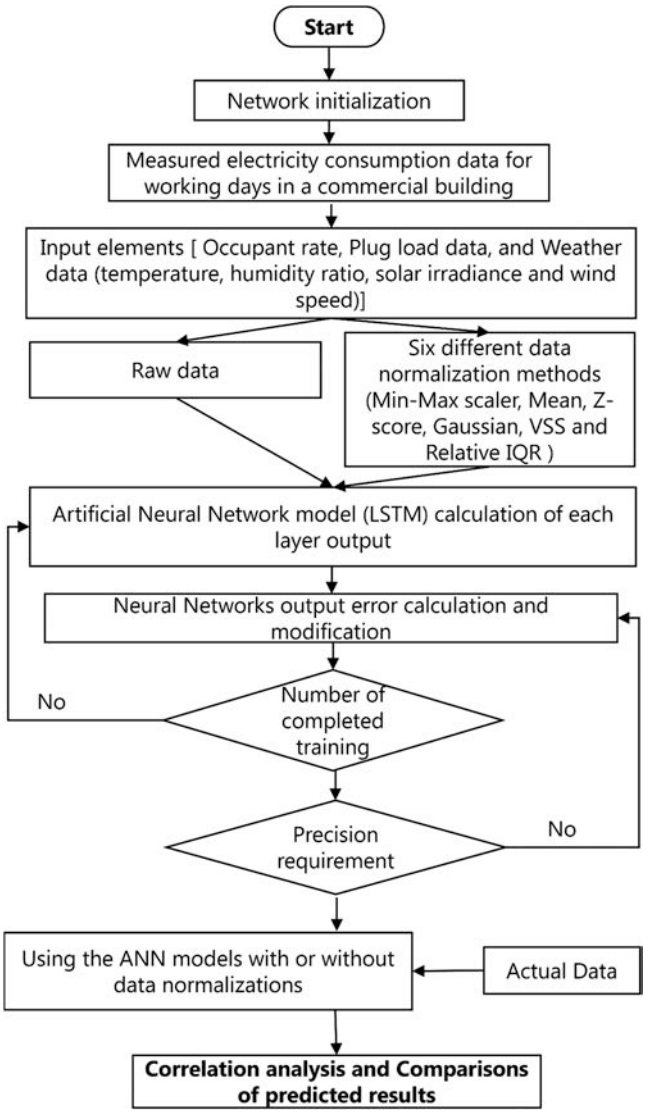

Fig. 2. The training procedure for the ANN methodology involving data normalization tools.

and specified in the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 14–2002 (ASHRAE, 2002):

$$CVRMSE(\%) = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2\right]^{1/2}}{\bar{y}} \times 100 \tag{14}$$

$$NMBE(\%) = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)}{N \times \bar{y}} \times 100 \tag{15}$$

The LSTM model undergoes individual training and testing with their respective normalization approaches. This methodology facilitates a thorough comparison evaluation of the analysis and accuracy of each ANN algorithm, targeting the particular objectives defined in the research. The study's objective is to assess how different normalization methods impact the accuracy of ANNs and analyze the correlation effectiveness of each element using ANNs.

In summary, Fig. 2 offers a succinct and informative visual summary, enhancing comprehension of the data-driven analysis setup and underscoring the importance of employing different normalization and correlation approaches in this study.

3.8. Gathered data and evaluation

The research employed data obtained from a commercial building located in the southern region of Philadelphia, Pennsylvania, USA. The structure spans an area of around 6000 square meters and consists of three stories and a bottom floor. The building designated 40 % of its area for offices and 60 % for shared facilities, while implementing a sub-metering system to monitor electricity consumption in different circuits. The dataset covered a period of 185 days, starting from August 2013, and ending on April 2014. It contained detailed information on building occupancy and electricity use, recorded on an hourly basis.

The U.S. National Solar Radiation Database (NSRDB) was utilized to include local weather data (Laboratory, 2023). was utilized for this numerical analysis. The overall dataset were classified into only working days, enabling a thorough investigation of how plug load usage, indoor occupant rate and surrounding environment impacted electricity usage in various situations. The data for the LSTM model simulation was split between 90 days for training and 14 days for testing. The choice to disregard data from non-working days, such as weekends and holidays, was made due to the significant disparities in energy consumption patterns seen during these timeframes in comparison to weekdays. The LSTM models utilized input nodes including plug load data, indoor occupant rate, dry bulb temperature, specific humidity, global horizontal irradiance (GHI), and wind speed in order to forecast the energy usage of the commercial building in working days.

4. Results

This study presented various methodologies for evaluating the performance of different six normalization methods and the correlation between plug load statistics, building occupancy rates, and local meteorological elements, and their influence on the real power usage of a commercial building over different seasons. The study utilized LSTM Neural Network algorithm to predict power usage patterns for a commercial building. The models underwent training using the dataset that included 185 business days or 4440 hours. The dataset incorporated many factors like as the building's plug load value (kW), indoor occupancy ratio, and meteorological characteristics including Global Horizontal Irradiation (GHI) (W/m2), wind speed (m/s), dry bulb temperature (°C), and specific humidity ratio (g/kg).

This research analyzed the precision and accuracy rates of the six data normalization methods, investigating the correlational impact factor of plug load, indoor occupancy ratio, and surrounding ambient meteorological conditions on the total electric usage in the building. Moreover, the utilization of LSTM technique facilitated the forecasting of extended-term electricity consumption by considering fluctuations in plug load, indoor occupant rates, and meteorological conditions. The models underwent training using 165 working days (equivalent to 3960 hours) of data. Their performance was assessed by comparing the anticipated total electricity consumption with the actual observed values, using 20 working days (equivalent to 480 hours) of test data.

Concisely, the study assessed six normalization models in combination with the LSTM method to forecast electrical energy usages in the commercial building on weekdays. All six normalization models exhibited precise predictions, however, the Min-Max, VSS, and Relative IQR normalization models displayed outstanding performance, showcasing superior accuracy (varying from 2.32 % to 18.44 %) and decreased error rates. Figs. 3 and 4 illustrate the predictive precision of the LSTM model, both when data normalization is used and when it is not. The results demonstrate a robust association between the projected values and the real values measured in both circumstances. The study's primary results highlight that the Min-Max, VSS, and Relative IQR normalization approaches surpass raw data and alternative normalizing strategies regarding prediction accuracy when implemented in the LSTM model. When combined with the LSTM approach, the Min-Max, VSS, and IQR scalers provide distinct benefits in enhancing prediction accuracy by reducing different error rates. By employing this normalizing technique, the forecasted interquartile range (IQR) is decreased, resulting in a narrower and better accurate prediction interval. Rescaling the data to a given range guarantees that the model's predictive performance is limited to a shorter interval, hence improving the accuracy of representing prediction uncertainty. In addition, the Min-Max and IQR scaler help reduce the greatest and least error rates between projected and real values. Normalization strategies aim to equalize

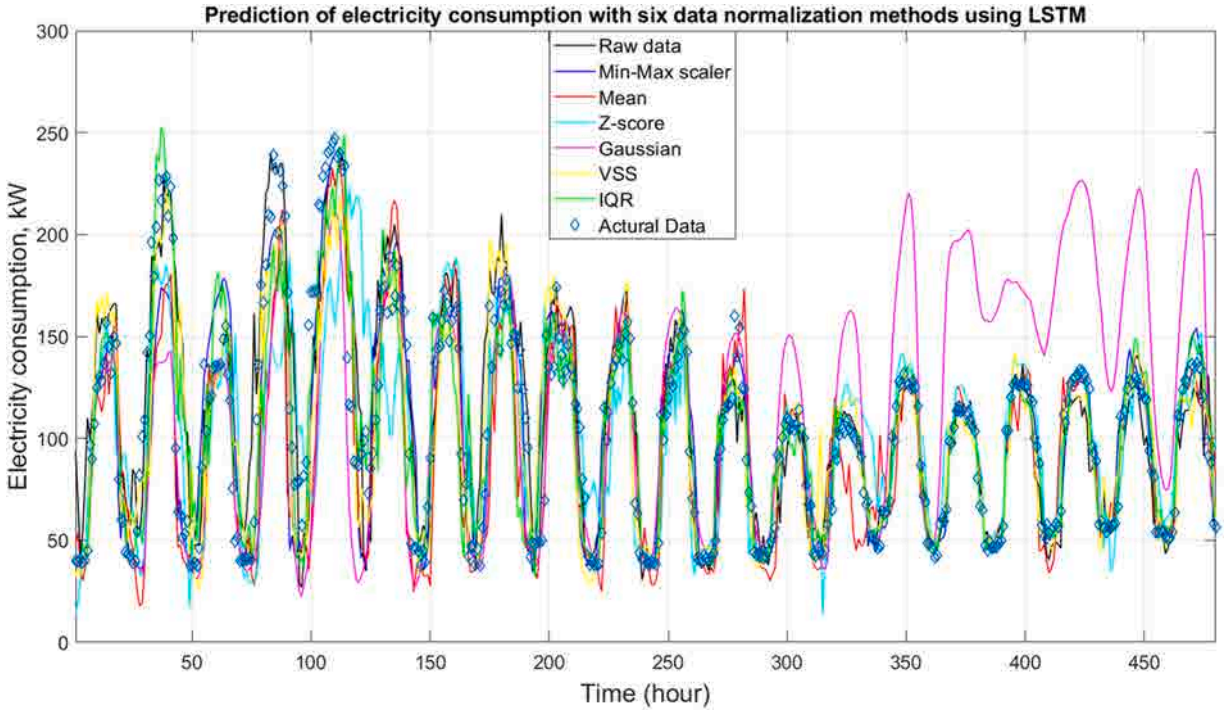

Fig. 3. Forecasting of electricity usages via the LSTM method in comparisons across six data normalization techniques, raw data and actual data.

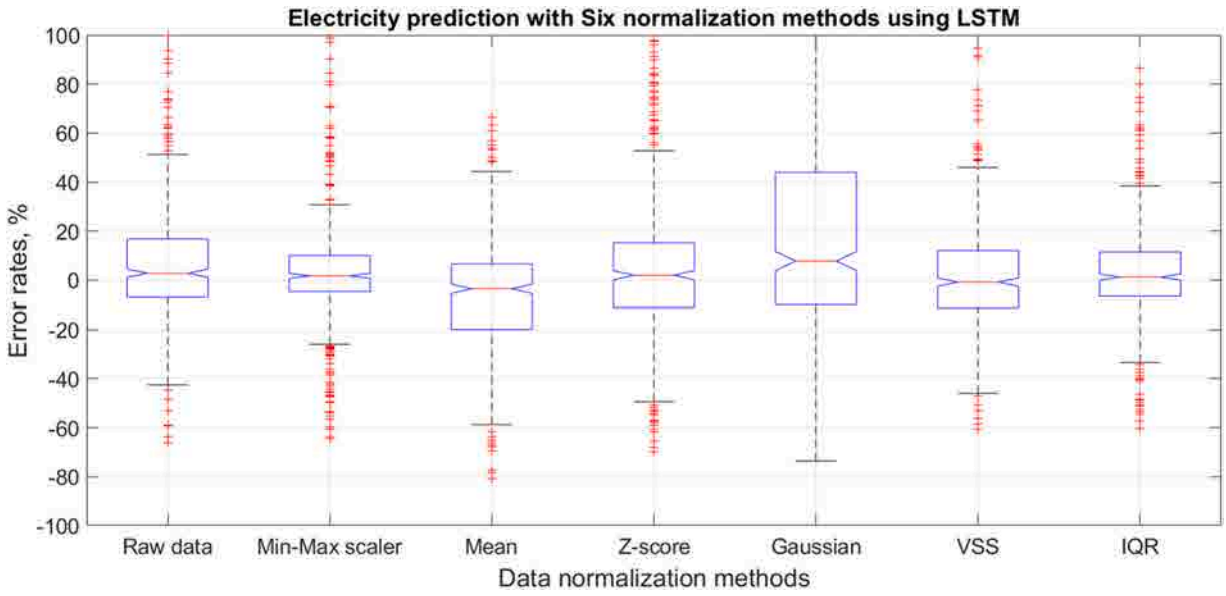

Fig. 4. Error rates of the LSTM algorithm utilizing six distinct normalization techniques and raw data.

feature weights and enhance model generalization, resulting in reduced error magnitudes and more uniform and equitable predictions across various value points.

Normalizing the data enables the LSTM method to effectively manage differing feature size, resulting in enhanced analysis performance in comparison to utilizing raw data. The overarching findings from the study emphasize the critical role of data normalization techniques, specifically Min-Max, VSS, and IQR, as a vital preprocessing step to augment the predictive abilities of the LSTM algorithm. In direct contrast to utilizing raw data, the incorporation of normalization methods confers a distinct advantage, showcasing the ability to generate predictions that are more accurate and reliable.

Table 1 outlines the outcomes of average error rate, CVRMSE, and NMBE for the LSTM algorithm when applied to raw data and six distinct normalization models. Among these models, VSS normalization yielded the lowest average error rate (3.30). However, the relative interquartile normalization (IQR) demonstrated superior performance with the least CVRMSE (3.26) and NMBE (-0.14) values, showcasing its effectiveness in comparison to alternative ANN methodologies. Additionally, Min-Max normalization exhibited commendable performance in predicting electricity consumption (CVRMSE: 15.77, NMBE: −0.72). Conversely, the LSTM metholdology with Mean and Gaussian normalization showed higher CVRMSE (180.08) and NMBE (8.22) values, as well as CVRMSE (295.04) and NMBE (-13.47) values in comparison to utilizing raw data value. Additionally the Z-score method also proved to be a viable option for predicting energy consumption using LSTM. Data normalization significantly influences the predictive accuracy of various ANN algorithms, with their efficacy contingent on the particular methodology and dataset types. These insights gleaned from Table 1 illustrates the benefits and drawbacks of employing these normalization methods to enhance prediction performance in energy consumption estimating across different normalization methods.

These results sought to assess the relationship among normalization techniques and these effects on power usage in an office building. The LSTM algorithm and six distinct normalization techniques were utilized to forecast electricity usage patterns. These results indicate that the LSTM method utilizing the normalization methods, Min-Max and IQR demonstrates superior prediction performance relative to alternative normalization techniques and the use of raw data. This study selected two effective data normalization method, Min-Max and IQR to analyze the correlation and impact each average input element.

Figs. 5 and 10 illustrate the correlation impact factor (CIF) of each average input element when its value is artificially increased and decreased by 5 %, affecting electricity consumption in a commercial structure. The CIF computation involves adjusting testing values by ± 5 % after ANN training to create artificial test data. This variation represents each parameter's influence, assuming a linear relationship. The modified outcomes are then compared to the actual results using the original test data as a reference. The CIF value is calculated by analyzing the magnitude of variation in the predicted data, providing a measure of each parameter's impact on overall electricity usage.

The findings are derived from LSTM prediction methods that incorporate two normalization techniques: Min-Max and IQR normalization, together with raw data. The examination considers seasonal variations, such as spring and fall, summer, and winter. Plug loads were the main drivers of electricity consumption in all seasons according to both normalization methods selected and raw data, as detected by LSTM algorithm. Temperature and humidity ratio had considerable effects on power use, especially during intermediate and summer seasons, compared to other climate parameters. Occupancy ratio affected energy

Table 1
Evaluation of LSTM performance utilizing raw data versus six data normalization techniques.

ANN method	Average error rates, %						
	Raw	Min-Max	Mean	Z-score	Gaussian	VSS	IQR
LSTM	10.21	3.91	−6.31	6.75	21.74	3.30	4.43
	CVRMSE, %						
	Raw	Min-Max	Mean	Z-score	Gaussian	VSS	IQR
LSTM	119.93	15.77	180.08	16.44	295.04	20.30	3.26
	NMBE, %						
	Raw	Min-Max	Mean	Z-score	Gaussian	VSS	IQR
LSTM	−5.47	−0.72	8.22	0.751	−13.47	0.93	−0.14

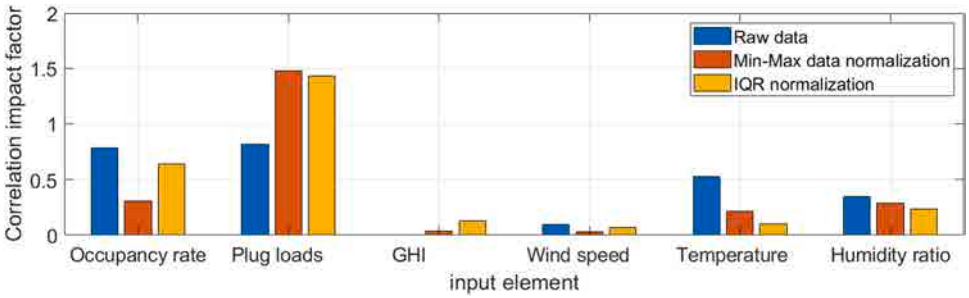

Fig. 5. The average correlation impact factors of various input elements after a 5 % increase in value during the spring and fall seasons, employing the LSTM algorithm with two normalization techniques and raw data.

usage in the intermediate season, whereas solar irradiance had a significant effect in the summer season. Wind speed has little impact on electricity use in any seasons. Comparing pairs of figures, such as Fig. 5 and Fig. 6, Fig. 7 and Fig. 8, and Fig. 9 and Fig. 10, these are presented that the correlation impact factor patterns exhibited similarities following a 5 % increase and decrease in value during seasonal changes, albeit at varying rates. The normalization methods played a crucial role, showcasing different values in correlation impact factors, with a more pronounced sensitivity observed after a 5 % decrease in value during the intermediate season. During the winter season, electricity consumption remained unaffected by climate variations, as the central heating system with a gas boiler mitigated the impact of cold climates. Notably, plug loads were the primary influencing factor on results during this season. Conversely, in the summer season, the use of all-air ventilation system and chiller for cooling and dehumidification played a significant role in electricity consumption. These systems were major contributors to actual electricity consumption during warmer months. Previous studies have indicated that plug loads are closely linked to occupancy patterns [62–64]. We project that the plug load mainly mirrors both the indoor occupant rate and overall electric usage in the commercial building. Consequently, the tangible influence of the occupant number is considered to be less significant than that of plug loads.

Moreover, the directional effect of the correlation impact factor values becomes apparent when analyzing a 5 % increase and decrease in value throughout seasonal variations with the LSTM model. These factors lack symmetry. A notable finding from this research is the capacity to predict how varied in these input parameters can impact the directions of energy usage in a building. It aids in determining which aspect can have a greater or lesser impact on the results. This study effectively demonstrated how data normalization approaches impact electricity usage forecasting with LSTM algorithm. However this study found some limitations that need to be investigated in future studies. The selection of a single commercial building for data gathering not adequately reflect the wide variety of building usages and commercial occupant patterns in the United States. Variations in meteorological circumstance, tenant pattern, plug loads, and thermal comfort may result in varying prediction and error rates when predicting energy usage for different types of buildings. Subsequent research should incorporate data from a broader range of construction types to improve the applicability of forecasting methodology using different standardization techniques. Furthermore, while the LSTM model used in this study showed promising outcomes, there are additional sophisticated Artificial Neural Network algorithms that can improve prediction accuracy even further. Studying the effectiveness of various Artificial Neural Network algorithms alongside different normalization techniques could provide useful insights for enhancing the accuracy of energy usage estimation jobs. Today advancements in normalization approaches should be examined to assess this influence on forecasting performance. Testing forecast models using data normalization techniques on external datasets or undertaking extended monitoring in actual energy usage situations may offer a better thorough assessment of the methodologies' analysis across various environment. However, constraint of the results is that the choice of a technique for converting input variables is strongly determined by the characteristics of those variables. Moreover, the procedure of converting unprocessed data may lead to a decrease in the quantity of information preserved from the initial dataset.

This research mainly contributed to applicating the influence of data normalization methods on predicting electricity usage utilizing LSTM model and six normalization methods. Future research can focus on addressing limitations and improving prediction accuracy and model generalizability.

5. Conclusion

This research emphasizes the innovative enhanced assessment techniques to analyze the effectiveness of six data normalization methodologies employing the LSTM algorithm on power usages forecasting in a typical commercial building in the United State America. The research examined the relationship between various data normalization techniques and their integration with LSTM algorithms for the electricity usages in the commercial building. The LSTM methodology was constructed utilizing input nodes from metrological data (solar irradiance, wind velocity, ambient dry bulb temperature, and specific humidity),

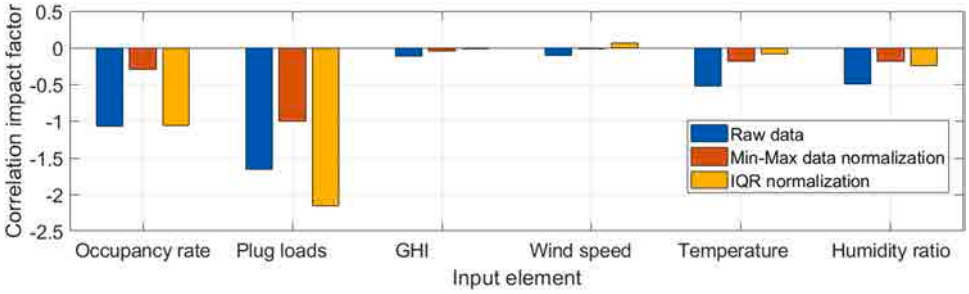

Fig. 6. The average correlation impact factors of various input elements after a 5 % decrease in value during the spring and fall seasons, employing the LSTM algorithm with two normalization techniques and raw data.

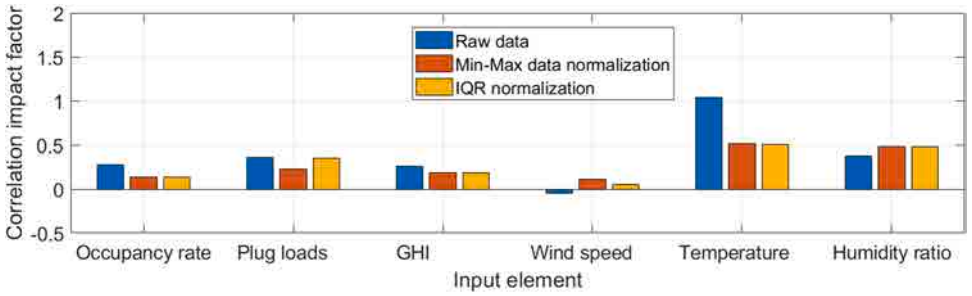

Fig. 7. The average correlation impact factors of various input elements after a 5 % increase in value during the summer seasons, employing the LSTM algorithm with two normalization techniques and raw data.

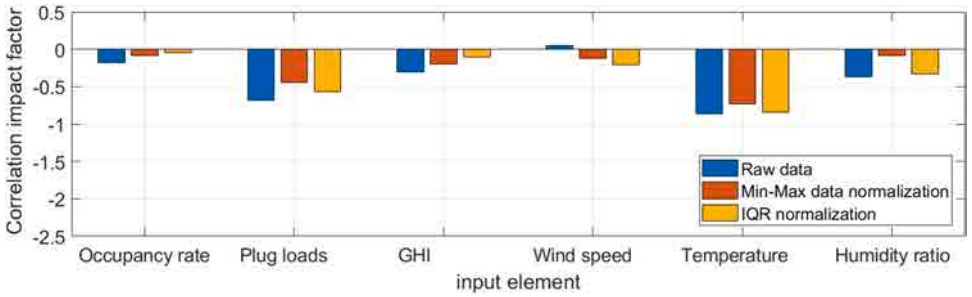

Fig. 8. The average correlation impact factors of various input elements after a 5 % decrease in value during the summer seasons, employing the LSTM algorithm with two normalization methods and raw data.

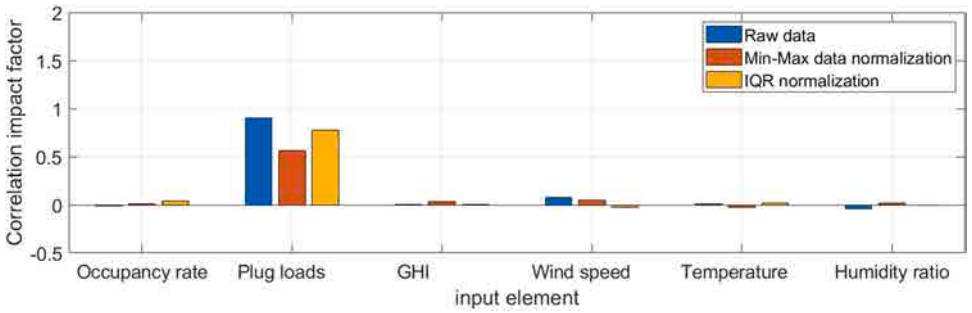

Fig. 9. The average correlation impact factors of various input elements after a 5 % increase in value during the winter seasons, employing the LSTM algorithm with two normalization techniques and raw data.

plug load data and occupant numbers. The model was subsequently assessed utilizing six different data standardization techniques: Minimum-Maximum, Mean value, Z-score, Gaussian normalization, VSS and IQR, over both a learning process and a test dataset. This research compares the performance of the LSTM algorithm using raw data and six normalization methods. VSS normalization achieved the lowest average error rate (3.30), while IQR normalization performed best overall with the lowest CVRMSE (3.26) and NMBE (-0.14). Min-Max normalization also showed strong results. In contrast, Mean and Gaussian normalizations resulted in significantly higher errors than using raw data. Data normalization techniques significantly impact this predictive analysis of the LSTM algorithm with its impact varying based on the particular

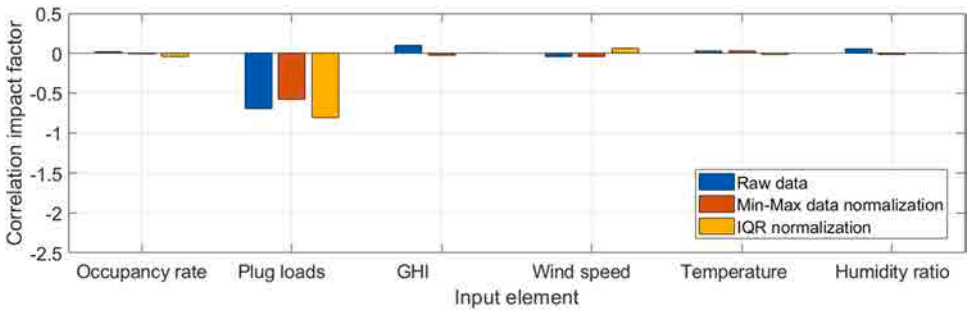

Fig. 10. The average correlation impact factors of various input elements after a 5 % decrease in value during the winter seasons, employing the LSTM algorithm with two normalization techniques and raw data.

methodology and dataset utilized. The results highlight the need of choosing suitable normalization techniques tailored to particular use and the balance between enhanced efficiency and computational difficulty.

This study represents the correlation impact factor each average input element which has a 5 % increased and decreased in value effectiveness on the usage of electricity in the commercial building using the LSTM algorithm with two effective normalization methods, Min-Max and IQR normalization as well as raw data. The assessments are conducted in the context of seasonal variations, including spring and fall, summer, and winter seasons. The plug load data mainly influenced on the total usage of electricity in two major normalization methods and raw data with LSTM algorithms across seasons, and both primary features: dry bulb temperature and specific humidity, exerted notable effect on the actual energy power usage in intermediate and summer seasons compared with other climates elements. The occupancy ratio had influences on the energy usage in intermediate season and the solar irradiance had specially impacted on summer season, however wind speed did rarely influence on the total energy consumption in all seasons. The correlation impact factor patterns presented similarities after a 5 % increase and decrease in value during the seasonal changes but not same rates. Depending on the normalization methods, the correlation impact factors presented different values, and the factors varied more sensitively after a 5 % decrease in value in intermediate season. The proposed the LSTM model and normalization techniques were effective in forecasting prolonged electricity usages and presenting correlation impact factors in different input elements in built environments. The results provide deeper understanding of building energy performance assessment and could be utilized to evaluate the effectiveness of various normalization techniques and artificial neural networks algorithms for particular building elements and dataset categories. Subsequent study may improve prediction accuracy by incorporating new input variables and investigating better optimal combinations of ANN algorithms with particular normalization techniques. To sum up, the research provides an extensive examination of the influence of normalization techniques on energy usage prediction using LSTM model, offering significant insights with enhanced analysis assessment using correlation impact values for further studies for prediction of building energy usage and evaluating of system performances. Nevertheless, the results lack the potential to be utilized across a wider dataset array. Hence, it is crucial to collect more data and evaluate performance using different algorithms and normalization techniques in order to broaden the scope of the assessment for future works.

CRediT authorship contribution statement

Sanghyuk Lee: Validation, Supervision, Formal analysis, Conceptualization. **Jelena Srebric:** Writing – review & editing, Investigation, Data curation. **Moon Keun Kim:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yang-Seon Kim:** Writing – original draft, Investigation, Data curation, Conceptualization. **Nuodi Fu:** Investigation, Formal analysis, Data curation. **Jiying Liu:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis. **Junqi Wang:** Validation, Methodology.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

References

ASHRAE. (2002). ASHRAE guideline 14–2002: Measurement of energy and demand savings. In: ASHRAE.

Barone, G., Buonomano, A., Forzano, C., Giuzio, G.F., Palombo, A., 2021. Increasing renewable energy penetration and energy independence of island communities: a novel dynamic simulation approach for energy, economic, and environmental analysis, and optimization. J. Clean. Prod. 311, 127558. https://doi.org/10.1016/j.jclepro.2021.127558.

Bunn, D.W., 2000. Forecasting loads and prices in competitive power markets. Proc. IEEE 88 (2), 163–169. https://doi.org/10.1109/5.823996.

Caro, E., Juan, J., Cara, J., 2020. Periodically correlated models for short-term electricity load forecasting. Appl. Math. Comput. 364, 124642. https://doi.org/10.1016/j.amc.2019.124642.

Chandra, Y.P., Matuska, T., 2022. Intelligent data systems for building energy workflow: data pipelines, LSTM efficiency prediction and more. Energy Build. 267, 112135. https://doi.org/10.1016/j.enbuild.2022.112135.

Cheng, Y.L., Lim, M.H., Hui, K.H., 2022. Impact of internet of things paradigm towards energy consumption prediction: a systematic literature review. Sustain. Cities Soc. 78, 103624. https://doi.org/10.1016/j.scs.2021.103624.

El Bouchefry, K., de Souza, R.S., 2020. Chapter 12 - Learning in Big Data: Introduction to Machine Learning. In: Škoda, P., Adam, F. (Eds.), Knowledge Discovery in Big Data from Astronomy and Earth Observation. Elsevier, pp. 225–249. https://doi.org/10.1016/B978-0-12-819154-5.00023-0.

Graves, A., 2012. Long Short-Term Memory. In: Graves, A. (Ed.), Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, pp. 37–45. https://doi.org/10.1007/978-3-642-24797-2_4.

Hasan, M.W., 2023. Building an IoT temperature and humidity forecasting model based on long short-term memory (LSTM) with improved whale optimization algorithm. Mem. - Mater., Devices, Circuits Syst. 6, 100086. https://doi.org/10.1016/j.memori.2023.100086.

Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., Amira, A., 2021. Artificial intelligence based anomaly detection of energy consumption in buildings: a review, current trends and new perspectives. Appl. Energy 287, 116601. https://doi.org/10.1016/j.apenergy.2021.116601.

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

Hou, S.L., Wei, J.H., Hou, M.L., Xu, J.Q., Han, L., 2025. A hydrological knowledge-informed LSTM model for monthly streamflow reconstruction using distributed data: application to typical rivers across the Tibetan plateau. J. Hydrol. 649, 132409. https://doi.org/10.1016/j.jhydrol.2024.132409.

Jinmiao, C., & Chaudhari, N.S. (2005, 31 July-4 Aug. 2005). Protein secondary structure prediction with a hybrid RNN/HMM system. Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.,

Kim, K.M., Cha, J., Lee, E., Pham, H.V., Lee, S., Theera-Umpon, N., 2019. Simplified neural network model design with sensitivity analysis and electricity consumption prediction in a commercial building. Energies 12 (7). https://doi.org/10.3390/en12071201.

Kim, M.K., Kim, Y.S., Srebric, J., 2020. Impact of correlation of plug load data, occupancy rates and local weather conditions on electricity consumption in a building using four back-propagation neural network models. Sustain. Cities Soc. 62, 102321. https://doi.org/10.1016/j.scs.2020.102321.

Kim, M.K., Kim, Y.-S., Srebric, J., 2020. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustain. Cities Soc. 62, 102385. https://doi.org/10.1016/j.scs.2020.102385.

Kim, T.-Y., Cho, S.-B., 2019. Predicting residential energy consumption using CNN-LSTM neural networks. Energy 182, 72–81. https://doi.org/10.1016/j.energy.2019.05.230.

Kim, Y.-S., Srebric, J., 2017. Impact of occupancy rates on the building electricity consumption in commercial buildings. Energy Build. 138, 591–600. https://doi.org/10.1016/j.enbuild.2016.12.056.

Kim, Y.-S., Kim, M.K., Fu, N., Liu, J., Wang, J., Srebric, J., 2024. Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. Sustain. Cities Soc., 105570 https://doi.org/10.1016/j.scs.2024.105570.

Kumar Dubey, A., Kumar, A., García-Díaz, V., Kumar Sharma, A., Kanhaiya, K., 2021. Study and analysis of SARIMA and LSTM in forecasting time series data. Sustain. Energy Technol. Assess. 47, 101474. https://doi.org/10.1016/j.seta.2021.101474.

Laboratory, T.N.R.E. (2023). *NSRDB: National Solar Radiation Database*. ⟨https://nsrdb.nrel.gov/⟩.

Li, Y., Tong, Z., Tong, S., Westerdahl, D., 2022. A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation. Sustain. Cities Soc. 76, 103481. https://doi.org/10.1016/j.scs.2021.103481.

Li, Y., Xiao, L.F., Wei, H.D., Li, D.Y., Li, X., 2025. A comparative study of LSTM and temporal convolutional network models for semisubmersible platform wave runup prediction. J. Offshore Mech. Arct. Eng. -Trans. Asme 147 (1), 011202. https://doi.org/10.1115/1.4063266.

Liu, X.Y., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X.B., Leung, K.S., Cheng, L.X., 2019. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front. Bioeng. Biotechnol. 7, 358. https://doi.org/10.3389/fbioe.2019.00358.

Lu, C.J., Li, S.H., Lu, Z.J., 2022. Building energy prediction using artificial neural networks: a literature survey. Energy Build. 262, 111718. https://doi.org/10.1016/j.enbuild.2021.111718.

Masood, A., Ahmad, K., 2021. A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: fundamentals, application and performance. J. Clean. Prod. 322, 129072. https://doi.org/10.1016/j.jclepro.2021.129072.
Peng, C., Tao, Y., Chen, Z., Zhang, Y., Sun, X., 2022. Multi-source transfer learning guided ensemble LSTM for building multi-load forecasting. Expert Syst. Appl. 202, 117194. https://doi.org/10.1016/j.eswa.2022.117194.
Peplinski, M., Dilkina, B., Chen, M., Silva, S.J., Ban-Weiss, G., Sanders, K.T., 2024. A machine learning framework to estimate residential electricity demand based on smart meter electricity, climate, building characteristics, and socioeconomic datasets. Appl. Energy 357, 122413. https://doi.org/10.1016/j.apenergy.2023.122413.
Ren, C., Cao, S.-J., 2020. Implementation and visualization of artificial intelligent ventilation control system using fast prediction models and limited monitoring data. Sustain. Cities Soc. 52, 101860. https://doi.org/10.1016/j.scs.2019.101860.
Rousseeuw, P.J., Hubert, M., 2011. Robust statistics for outlier detection. WIREs Data Min. Knowl. Discov. 1 (1), 73–79. https://doi.org/10.1002/widm.2.
Runge, J., Zmeureanu, R., 2019. Forecasting Energy Use in Buildings Using Artificial Neural Networks: A Review. Energies 12 (17), 3254. https://doi.org/10.3390/en12173254.
Saha, B., & Srivastava, D. (2014, 31 March-4 April 2014). Data quality: The other face of Big Data. 2014 IEEE 30th International Conference on Data Engineering,
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
da Silva, D.G., Meneses, A.A. d M., 2023. Comparing long short-term memory (LSTM) and bidirectional LSTM deep neural networks for power consumption prediction. Energy Rep. 10, 3315–3334. https://doi.org/10.1016/j.egyr.2023.09.175.
Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524.
Singh, D., Singh, B., 2022. Feature wise normalization: an effective way of normalizing data. Pattern Recognit. 122, 108307. https://doi.org/10.1016/j.patcog.2021.108307.
Singla, P., Duhan, M., Saroha, S., 2022. 10 - Different normalization techniques as data preprocessing for one step ahead forecasting of solar global horizontal irradiance. In: Dubey, A.K., Narang, S.K., Srivastav, A.L., Kumar, A., García-Díaz, V. (Eds.), Artificial Intelligence for Renewable Energy Systems. Woodhead Publishing, pp. 209–230. https://doi.org/10.1016/B978-0-323-90396-7.00004-3.
Soori, M., Arezoo, B., Dastres, R., 2023. Artificial neural networks in supply chain management, a review. J. Econ. Technol. 1, 179–196. https://doi.org/10.1016/j.ject.2023.11.002.
Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., Wei, X.T., 2023. A review of normalization and differential abundance methods for microbiome counts data. Wiley Interdiscip. Rev. -Comput. Stat. 15 (1), e1586. https://doi.org/10.1002/wics.1586.
Wang, H., Wang, J., Feng, Z., Haghighat, F., Cao, S.-J., 2023. An intelligent anti-infection ventilation strategy: from occupant-centric control and computer vision perspectives. Energy Build. 296, 113403. https://doi.org/10.1016/j.enbuild.2023.113403.
Wang, L.Y., Que, H.B., Wu, F., 2025. The CNN-LSTM-attention model for short term prediction of the polar motion. Meas. Sci. Technol. 36 (1), 016323. https://doi.org/10.1088/1361-6501/ad8be5.
Wang, Z.Y., Srinivasan, R.S., 2017. A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models. Renew. Sustain. Energy Rev. 75, 796–808. https://doi.org/10.1016/j.rser.2016.10.079.
Weerakody, P.B., Wong, K.W., Wang, G., Ela, W., 2021. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing 441, 161–178. https://doi.org/10.1016/j.neucom.2021.02.046.
Wu, Y.Z., Tao, R., Zhu, D., Xiao, R.F., 2025. Analysis and prediction of force characteristics of tubular turbine based on Hankel-DMD-LSTM. Eng. Appl. Comput. Fluid Mech. 19 (1), 2443122. https://doi.org/10.1080/19942060.2024.2443122.
Ye, Z., Kim, M.K., 2018. Predicting electricity consumption in a building using an optimized back-propagation and Levenberg–Marquardt back-propagation neural network: case study of a shopping mall in China. Sustain. Cities Soc. 42, 176–183. https://doi.org/10.1016/j.scs.2018.05.050.
Zhang, C., Ma, L., Luo, Z., Han, X., Zhao, T., 2024. Forecasting building plug load electricity consumption employing occupant-building interaction input features and bidirectional LSTM with improved swarm intelligent algorithms. Energy 288, 129651. https://doi.org/10.1016/j.energy.2023.129651.
Zhang, Z.H., Cheng, Y., Liu, N.C., 2014. Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. Scientometrics 101 (3), 1679–1693. https://doi.org/10.1007/s11192-014-1294-7.
Zhu, H.-C., Ren, C., Cao, S.-J., 2022. Dynamic sensing and control system using artificial intelligent techniques for non-uniform indoor environment. Build. Environ. 226, 109702. https://doi.org/10.1016/j.buildenv.2022.109702.