

Applications of Machine Learning in Green Building Data Processing

Zhenlin Wang

College of Greater Bay Area Design and Arts, South China Agricultural University, Guangzhou, China
jwongzlll@163.com

Abstract. Against the backdrop of the global "dual-carbon" goals driving the transformation of the building sector toward life-cycle green buildings, green buildings urgently require efficient, data-driven technologies. At the same time, traditional approaches face difficulties in processing multi-source heterogeneous building data and achieving dynamic optimization. This paper examines the utilization of machine learning in the processing of green building data through a literature review, focusing on the comprehensive stages of data preprocessing, feature extraction, data analysis, and building energy optimization, while assessing its applicability, technical methodologies, effectiveness, and existing limitations. The results indicate that appropriately applying machine learning to data processing tailored to the characteristics of buildings in different regions can significantly improve data utilization and prediction accuracy, thereby effectively reducing building energy consumption and carbon emissions. However, challenges remain, including uneven data quality, insufficient model interpretability, and weak scenario adaptability. Future research should promote multimodal data fusion and the development of lightweight models to support the intelligent development of green buildings.

Keywords: Machine Learning, Green Buildings, Building Energy Consumption, Data Processing, Sustainable Buildings

1. Introduction

With social development and progress, awareness of environmental protection has increased, and ecological civilization has become a global consensus. Against this backdrop, the building sector, as a major source of resource consumption and carbon emissions, plays a critical role in green the transition toward sustainability.

In China, carbon emissions from building and infrastructure construction reached 2.78 billion tCO₂ in 2024. Carbon emissions from civil building operations amounted to 2.47 billion tCO₂, accounting for 22.1% of national energy consumption, while building operational energy consumption reached 1.30 billion tce, representing 21.8% of total energy consumption [1]. Globally, buildings account for 32% of energy use and 34% of carbon dioxide emissions [2]. These figures highlight the building sector's significant contribution to carbon emissions and underscore the urgency of promoting green, low-carbon development. However, traditional statistical analyses and

physics-based models often struggle to handle the diversity and complexity of data generated by green buildings, thereby limiting the accuracy of predictions and controls.

In recent years, advances in artificial intelligence have driven the application of machine learning in the building sector [3], creating new opportunities for green building development. Studies have shown that more accurate prediction of building energy systems can reduce building energy consumption by 10%–30% [3, 4]. Therefore, this paper systematically investigates the application of machine learning to the processing of green building energy-consumption data. By reviewing machine learning-based approaches and analyzing existing challenges, this study establishes a data-driven research framework following the sequence of data collection and preprocessing, energy-consumption feature analysis, and energy-consumption decision-making and optimization. It further summarizes the applicability of commonly used machine learning methods to different building types and climatic conditions, providing theoretical support for the optimization and management of green building energy-consumption data processing.

2. Theoretical foundations and current status of energy-consumption data processing

2.1. Fundamental theories of machine learning-based data processing and green building energy consumption

Machine learning learns from building energy-consumption data and identifies underlying patterns through algorithmic models. Its core application framework consists of three layers: the data input layer, the model training layer, and the result output layer. The input layer receives multi-source energy-consumption data generated across the green building life cycle, including real-time sensor data, operational data, and climatic data. The training layer conducts data mining and pattern recognition via computational models, whilst the output layer produces analytical results and optimisation suggestions [3, 5]. One of its key advantages is that it does not require reprogramming across different datasets [5], enabling the maximization of energy efficiency and enhancing indoor sustainability [6]. Moreover, machine learning can effectively process high-dimensional and heterogeneous life-cycle energy-consumption data, overcoming the limitations of traditional data-processing methods and providing data-driven support for green building energy management.

2.2. Characteristics of green building energy-consumption data and challenges of traditional processing methods

Green building energy-consumption data are influenced by multiple factors, including but not limited to physical, environmental, and social characteristics [7]. Consequently, such data exhibit features such as missing values, variability, disorder, complexity, and randomness. Conventional physics-based models depend heavily on comprehensive data inputs; thus, when data are deficient, these models fail to perform effectively, leading to diminished predictive reliability [7]. In addition, traditional models are generally limited to relatively simple data analysis and are unable to address the diversity and complexity of contemporary energy-consumption data [7]. In contrast, machine learning focuses primarily on the model training layer, thereby reducing constraints on input data while continuously improving inference and decision-making based on output results [5, 7].

3. Key application of machine learning in green building energy-consumption data processing

3.1. Collection and organization of life-cycle energy-consumption data for green buildings

The collection of life-cycle energy-consumption data for green buildings primarily focuses on three key stages: design, construction, and operation. Various methods are available for collecting energy-consumption data, including the use of Internet of Things (IoT) sensors and their integration with Building Information Modeling (BIM) technology, both of which enable large-scale acquisition of energy-consumption data [8]. In addition, Wireless Sensor Networks (WSNs) supported by big data technologies can facilitate energy-consumption monitoring at larger spatial scales [9]. Furthermore, the increasing adoption of wearable devices has further expanded the scope of energy-consumption data collection. After data acquisition, preliminary data visualization is typically conducted before processing, using histograms, density plots, pair plots, heat maps, and other visualization techniques to provide an initial overview of the data.

With the development of intelligent devices, an increasing amount of raw data has become available, providing sufficient support for the subsequent application of machine learning models.

3.2. Machine learning-based energy-consumption data processing

Before applying machine learning to energy-consumption data, the data must undergo preprocessing to facilitate the training and learning of models. Preprocessing workflow typically includes data screening, outlier removal, data denoising, normalization, and standardization.

Common methods used in this process are as follows:

(1) K-Means-Based Anomaly Detection and Removal: This is a typical unsupervised machine learning method that partitions samples into several non-overlapping subsets. Based on the user-defined value of k , it employs Euclidean distance as a measure of sample similarity and assigns data points to the nearest cluster [10]. This approach can efficiently identify anomalous patterns and outliers in large-scale energy-consumption datasets.

$$d = \sqrt{\{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2\}} \quad (1)$$

The K-Means clustering algorithm can effectively identify anomalous patterns and outliers in large-scale datasets of building energy consumption. It is suitable for building scenarios with significant fluctuations in energy consumption, such as office buildings, commercial complexes, and high-density residential areas, particularly for analyzing energy-consumption data strongly influenced by occupant mobility, equipment operating conditions, and climatic factors. Because the energy-consumption data from such buildings are typically characterized by large volumes, pronounced temporal fluctuations, and complex load characteristics, this method can efficiently distinguish normal from abnormal energy consumption. This provides a reliable data foundation for subsequent data cleaning, anomaly detection, and energy-consumption prediction.

(2) Artificial Neural Network (ANN): Artificial Neural Networks (ANNs) are typical supervised machine learning algorithms that establish nonlinear mapping relationships between input variables and target variables through learning from historical data features, thereby enabling the prediction and fitting of building energy-consumption data [11]. This method is also applied in subsequent

analyses of energy consumption data. As its underlying principles and applicable scenarios are similar, they are not discussed in detail here.

(3)Isolation Forest(iForest): Isolation Forest (iForest) is widely applied to outlier detection and data cleaning in building energy-consumption datasets. The algorithm constructs isolation trees by randomly partitioning the feature space. It identifies anomalies by noting that abnormal samples are more likely to be isolated in fewer partitioning steps. Compared with traditional distance-based or density-based anomaly detection methods, iForest does not require predefined data distributions or complex parameter tuning, thereby offering high computational efficiency, strong scalability, and strong engineering applicability [11].

This method is particularly suitable for building energy-consumption analysis in small- and medium-sized buildings and regions with relatively stable climatic conditions. Limited sample sizes, pronounced data fluctuations, and frequent periodic anomalies are typically characteristic of such buildings. Under these conditions, iForest can reliably identify local anomalous data and outliers, thereby improving data quality and the stability of prediction models. In recent years, researchers have proposed extended methods, such as the Deep Isolation Forest (DIF) and the Functional Isolation Forest (FIF). By incorporating nonlinear feature mapping and function-space extension, these approaches enhance the ability to handle complex anomalous data and improve model adaptability to complex building operational data [11, 12].

(4)Feature normalization: Feature normalization is not, in essence, a machine learning algorithm; however, when combined with machine learning algorithms, it can reduce bias. Its primary function is to transform feature values into a consistent range [6]. A commonly used method is Z-score standardization. Consequently, its applicability is relatively constrained, as it is mainly used in conjunction with machine learning algorithms and is generally unaffected by differences in building types or climatic regions, thereby enabling a broader range of applications.

$$x' = \frac{x-\mu}{\sigma} \quad (2)$$

Overall, different machine learning algorithms demonstrate targeted applicability and practical effectiveness in processing green building energy-consumption data. They provide effective approaches for addressing the challenges posed by multi-source heterogeneous data and establish a reliable data foundation for subsequent energy-consumption analysis, prediction, and decision-making tasks based on machine learning.

3.3. Analysis and application of different machine learning methods for green building energy-consumption data characteristics

Machine learning algorithms are employed to mine features from energy-consumption data across different stages of the building life cycle, thereby identifying energy-consumption patterns and key influencing factors at each stage [13]. For different building types and climatic conditions, appropriate machine learning algorithms can be applied to identify and analyze critical energy-saving parameters and high-energy-consuming processes during building operation, providing targeted data analysis support for subsequent energy consumption prediction and optimization.

(1)Random Forest Regressor

Random Forest Regressor (RFR) is a supervised machine learning algorithm based on ensemble learning. During training, multiple decision trees are constructed, and the final prediction is obtained

by averaging their outputs. The algorithm randomly selects a subset of energy-consumption data features for learning during model training, thereby reducing the risk of overfitting associated with a single decision tree and improving the model's generalization capability and prediction accuracy [6].

$$\hat{X} = \frac{1}{M} \sum_{i=1}^M T_i(y) \quad (3)$$

In this formula, $\hat{X}$ denotes the target variable's predicted value, the input feature's forecasted value by M , and the prediction generated by the i th decision tree for the input feature is denoted by $T_i(y)$ [6].

The Random Forest Regressor is suitable for analyzing energy consumption in scenarios involving large-scale datasets and complex energy consumption characteristics, such as high-rise office buildings and commercial complexes. These buildings are typically influenced by multiple factors, including climatic conditions, occupant mobility, equipment operating status, and functional building use, resulting in energy-consumption data that are high-dimensional, highly nonlinear, and subject to significant fluctuations. Owing to its strong randomness and nonlinear fitting capabilities, the Random Forest Regressor can accurately predict building energy consumption and perform feature analysis. In addition, the algorithm demonstrates strong adaptability to different data distributions and climatic conditions, making it applicable to building energy-consumption studies across various climate zones and providing a broad range of practical applications.

(2) Gradient Boosting Regressors

Gradient Boosting Regressor (GBR) is a representative technique in ensemble learning. It generates a sequence of weak learners by iteratively correcting the errors of preceding learners and subsequently integrates their outputs to produce the final prediction. Consequently, the learning rate plays a critical role in determining the model's convergence behavior and predictive accuracy [6].

$$\hat{X} = \frac{1}{M} \sum_{i=1}^M \gamma_i T_i(y) \quad (4)$$

In this formula, $\hat{X}$ denotes the target variable's predicted value, M is the total number of poor learners in the group. The decreasing parameter or learning rate for the i -th weak learner is denoted by γ_i [6].

Building upon this framework, LightGBM Regressor and CatBoost Regressor have demonstrated strong capabilities in efficiently processing energy-consumption data and are particularly suitable for large-scale datasets. LightGBM constructs decision trees using a leaf-wise growth strategy rather than the conventional level-wise approach, resulting in higher computational efficiency, fewer nodes, and enhanced capability for deep data analysis and accurate prediction [6]. CatBoost Regressor performs continuous numerical prediction by partitioning the feature space into multiple smaller regions, each associated with a corresponding prediction value. Through iterative boosting, additional decision trees are continuously introduced to correct the errors of preceding models and further improve prediction accuracy [6].

Gradient boosting-based models are ideally suited for analysing energy consumption in factories, high-rise office buildings, commercial complexes, and other public structures that exhibit diverse energy-use patterns and substantial data volumes, particularly in areas with significant climatic variability and extreme weather conditions. When dealing with large-scale and highly fluctuating

datasets, gradient boosting regressors can effectively reduce prediction errors and improve predictive performance. As advanced variants of traditional gradient boosting algorithms, LightGBM and CatBoost offer superior computational efficiency and prediction accuracy for large datasets, making them particularly advantageous for large public buildings and buildings located in regions with extreme climatic conditions compared with smaller-scale buildings.

(3) K-Nearest Neighbors Regressor

K-Nearest Neighbors Regressor (KNN Regressor) predicts target values by aggregating the target values of the nearest neighboring data points within a specified feature space. Distance metrics, such as Euclidean distance, are typically used to calculate the distance between a target data point and the existing dataset. During prediction, the algorithm retrieves neighboring samples to form a local sample region. When a new data point is introduced, the average of the k nearest neighbors is calculated as the expected prediction value [14].

Its simplicity and strong predictive accuracy characterizes this algorithm; nevertheless, it is highly reliant on the quality and distribution of the input data. KNN, as a passive learning approach, is typically less appropriate for intricate application scenarios [12]. These characteristics make it particularly appropriate for energy-consumption analysis in small residential buildings, villas, and community-scale buildings. It is also well-suited to regions with mild climates, such as those with oceanic and Mediterranean climates. Because energy-consumption datasets in these buildings and regions are typically smaller in scale and less complex, the KNN model can provide relatively accurate predictions and analysis under such conditions.

(4) Artificial Neural Networks

Artificial Neural Networks (ANNs) are a typical supervised machine learning algorithm that can learn nonlinear relationships between input and target variables from historical data, thereby enabling the prediction and fitting of building energy-consumption data [13]. A typical ANN consists of an input layer, one or more hidden layers, and an output layer. By iteratively optimizing connection weights and continuously correcting prediction errors, the network learns complex data patterns. Due to the nonlinear feature-extraction capabilities of multilayer neurons, artificial neural networks may efficiently discern potential relationships within high-dimensional, temporal, and heterogeneous datasets, as well as predict, fit, and reconstruct energy consumption data. As a result, they are widely used in studies of building energy consumption [15].

This method is not only suitable for addressing data-missing issues in existing buildings caused by equipment failures, sensor errors, or discontinuous data collection, but also for addressing data-missing issues in new buildings. It can also be applied to energy-consumption prediction and performance evaluation during the building design stage. ANNs are particularly suitable for energy-consumption analysis in small- and medium-sized buildings and regions with relatively uniform climatic conditions. In these situations, passenger movement, equipment initiation and cessation activities, and ambient variables often produce intricate nonlinear features in energy consumption data, complicating the ability of conventional linear models to characterize their fundamental patterns. For high-energy-consumption green buildings located in hot-arid or hot-humid regions, ANNs can identify deep nonlinear and temporal relationships between building energy consumption and environmental variables such as air temperature, solar radiation, and humidity, thereby improving the accuracy of building energy-consumption analysis and prediction.

(5) Support Vector Machine & Naive Bayes/Bayes Classifiers

Support Vector Machine (SVM) is based on the principle of maximum-margin classification, in which data are separated by a hyperplane, providing strong generalization capability. The performance of SVM largely depends on the kernel function employed during model development.

Therefore, selecting an appropriate kernel function is essential for achieving satisfactory performance, although it may also increase computational time and complexity [16, 17].

Due to its computational expense, SVM is more appropriate for small-scale structures or those with specific functional requirements, as well as for areas with reasonably stable environmental conditions. In such situations, the comparatively limited data or specific application needs can mitigate the drawbacks of SVM. Due to its robust generalization capacity, SVM can effectively identify nonlinear correlations between climatic variables and energy use. Moreover, its superior capacity to interpret high-dimensional data on building design features renders it particularly adept at analyzing specialized green building datasets.

Bayesian classifiers are a class of probabilistic classification algorithms based on Bayes' theorem. Naive Bayes, the most commonly used subclass and the algorithm used in this study, simplifies computation by assuming conditional independence among features. It determines sample categories by inferring posterior probabilities from prior probabilities and can be implemented through Bayesian network structures. This method is computationally efficient and demonstrates strong tolerance to small-sample and missing-data scenarios [16].

Bayesian classifiers are relevant to various types of green buildings across diverse climatic regions, particularly for commercial complexes, hotels, and large public buildings that experience significant variations in occupant density, as well as those situated in hot-humid and hot-summer/cold-winter climate zones. The algorithm enables rapid classification and is well-suited to scenarios with missing monitoring data.

Compared with Bayesian classifiers, SVMs generally incur higher computational costs but demonstrate stronger performance on high-dimensional data. In contrast, Bayesian classifiers are more computationally efficient and more tolerant of missing data. Compared with Random Forest, both methods possess strong generalization ability; however, Bayesian classifiers are more suitable for small-sample scenarios, whereas SVMs are better suited to high-dimensional datasets.

(6) Reinforcement Learning & Semi-Supervised Learning

Reinforcement Learning (RL) is a dynamic learning paradigm that does not require pre-labeled information regarding the distance between a current state and a target outcome. Essentially, it is a learning framework based on a trial-and-error approach, in which an agent persistently investigates and selects behavioral patterns that produce superior results and advance toward the intended goal. Reinforcement learning integrates features of both supervised and unsupervised learning. The learning process often commences with random exploration and subsequently enhances decision-making processes through positive reinforcement and iterative optimization [16].

This approach is particularly effective for energy-optimization problems characterized by strong dynamics and real-time control requirements. Buildings located in regions with highly variable climatic conditions, as well as zero-carbon or near-zero-carbon buildings, and intelligent buildings that aim to minimize energy consumption while maximizing renewable energy utilization, all rely on dynamic operational strategies. Similarly, industrial facilities often require real-time equipment regulation in response to changing production demands. These application scenarios closely align with the characteristics and advantages of reinforcement learning.

Semi-Supervised Learning (SSL) integrates the paradigms of supervised and unsupervised learning by training models with a small amount of labeled data and a large amount of unlabeled data. By exploiting the intrinsic distribution patterns within datasets, SSL addresses the challenges of limited labeled samples and high labeling costs, thereby improving model generalization and enabling both regression and classification tasks.

This approach is particularly suitable for building energy-consumption prediction and pattern recognition in data-scarce scenarios, such as rural or remote regions. In these areas, factors such as limited transportation accessibility, underdeveloped infrastructure, and constrained data-collection conditions often result in short monitoring periods and substantial data deficiencies. By uncovering energy-consumption characteristics and patterns from limited datasets, semi-supervised learning can provide valuable support for the widespread design of green buildings and the development of region-specific energy-efficiency optimization strategies.

3.4. Machine learning-based energy consumption decision-making and optimization

Based on energy-consumption data analysis and predictions at different stages, targeted energy-optimization strategies can be formulated. These include optimizing building layouts and material selection during the planning and design stage, improving resource allocation and construction processes during the construction stage, optimizing equipment parameters during the operation stage, and enhancing waste sorting and resource reuse strategies during the demolition and recycling stage, thereby minimizing energy consumption throughout the building life cycle.

The proposed energy-optimization strategies should undergo engineering feasibility analysis. Evaluating the operability and implementation efficacy of methods can be achieved by examining practical green building scenarios, cost budgets, and technical conditions. Anticipate potential obstacles during implementation and offer related mitigation measures to ensure the successful execution of optimisation strategies and the attainment of building energy-saving objectives.

4. Conclusion

This study addresses the need for energy-consumption data processing throughout the life cycle of green buildings. It systematically summarizes the applications and adaptation strategies of different machine learning algorithms for energy-consumption data preprocessing across various building types and climatic regions. The study clarifies the applicable scenarios and implementation considerations for algorithms such as K-Means clustering, Artificial Neural Networks (ANNs), and Isolation Forest, providing clear guidance and references for selecting and applying machine learning algorithms to green building energy-consumption data processing across different scenarios. It also facilitates a deeper understanding of algorithm characteristics and supports the appropriate selection of models based on engineering requirements.

Due to space limitations, this study does not provide an in-depth analysis of the theories of machine learning algorithms, nor does it include a comparative investigation of more advanced and diverse algorithmic models. Future research may build upon this work by conducting comparative evaluations of different machine learning algorithms and exploring multi-algorithm integrated modeling approaches. In addition, refined empirical analyses based on the energy-consumption characteristics of different climatic regions and building types should be conducted to further enrich the application framework for machine learning in green building energy-consumption data processing and to improve the accuracy and practical applicability of algorithm deployment.

References

- [1] China Building Energy Efficiency Association, Chongqing University. (2026). *Research report on carbon emissions in China's urban and rural construction field (2025)*. China Building Energy Efficiency Association.
- [2] UNEP, GlobalABC. (2025). *Global status report for buildings and construction 2024/2025*. United Nations Environment Programme.

- [3] Olu-Ajayi, R., Alaka, H., Sulaimon, I., et al. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, 45, 103406.
- [4] Colmenar-Santos, A., de Lober, L. N. T., Borge-Diez, D., et al. (2013). Solutions to reduce energy consumption in the management of large buildings. *Energy and Buildings*, 56, 66-77.
- [5] Balali, F., Nouri, J., Nasiri, A., & Zhao, T. (2020). Machine learning principles. In *Data intensive industrial asset management*. Springer. https://doi.org/10.1007/978-3-030-35930-0_8
- [6] Mahmood, S., Sun, H., Ali Alhussan, A., et al. (2024). Active learning-based machine learning approach for enhancing environmental sustainability in green building energy consumption. *Sci Rep*, 14, 19894. <https://doi.org/10.1038/s41598-024-70729-4>
- [7] Pan, Y., & Zhang, L. (2020). Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy*, 268, 114965. <https://doi.org/10.1016/j.apenergy.2020.114965>
- [8] Guerra-Santin, O., & Tweed, C. A. (2015). In-use monitoring of buildings: An overview of data collection methods. *Energy and Buildings*, 93, 189-207.
- [9] Plageras, A. P., Psannis, K. E., Stergiou, C., Wang, H., & Gupta, B. B. (2018). Efficient IoT-based sensor BIG data collection–processing and analysis in smart buildings. *Future Generation Computer Systems*, 82, 349-357.
- [10] Zhang, J., Zhang, H., Ding, S., & Zhang, X. (2021). Power consumption predicting and anomaly detection based on transformer and K-means. *Frontiers in Energy Research*, 9, 779587. <https://doi.org/10.3389/fenrg.2021.779587>
- [11] Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12591-12604.
- [12] Staerman, G., Mozharovskiy, P., Cléménçon, S., & d'Alché-Buc, F. (2019, October). Functional isolation forest. In *Asian Conference on Machine Learning* (pp. 332-347). PMLR.
- [13] Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: A review. *Visualization in Engineering*, 6(1), 5.
- [14] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer Berlin Heidelberg.
- [15] Islam, M., Chen, G., & Jin, S. (2019). An overview of neural network. *American Journal of Neural Networks and Applications*, 5(1), 7-11.
- [16] Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine learning algorithms: A review. *Mach. Learn*, 6(2), 916-922.
- [17] Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. *Qubahan Academic Journal*, 1(2), 81-90.