energies

MDPI

Systematic Review

Data Preprocessing Techniques for Machine Learning Towards Improving Building Energy Performance: A Systematic Review

Weixian Mu, Riccardo Cardelli and Simone Ferrari *

Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Via Ponzio 31, 20133 Milan, Italy; weixian.mu@polimi.it (W.M.); riccardo.cardelli@polimi.it (R.C.)
* Correspondence: simone.ferrari@polimi.it

Abstract

Enhancing building energy performance has become an essential goal, particularly as building energy management systems (BEMSs) increasingly rely on high-quality data and reliable predictive models. Although machine learning (ML) models have been widely applied to building energy prediction, optimisation, and management, their reliability in practice is often constrained by data preprocessing rather than algorithm selection. Existing studies often emphasise algorithmic development while providing limited systematic investigation of preprocessing practices, leading to methodological misconceptions and reduced robustness of ML-driven building energy management. As a novel contribution, this article presents a systematic review of 73 scientific articles published from 2020 to 2025 in the field of preprocessing practices. To this goal, a three-step data preprocessing workflow is organised, comprising data analysis, data preparation, and feature engineering. The strengths, limitations, and recurring misconceptions of preprocessing techniques adopted in the analysed studies are synthesised, with emphasis on their impact on prediction accuracy, interpretability, and model robustness. As a result, this review reframes the data preprocessing stage as a decision-making process in which data analysis and the energy improvement task constrain and inform subsequent data preparation and feature engineering steps to address building energy performance enhancement tasks.

Keywords: building energy performance; building energy management systems; machine learning; data preprocessing; systematic review

Academic Editors: José Carlos Magalhães Pires, Eugenio Meloni, Iva Ridjan Skov, Giorgio Vilardi, Antonio Zuorro, Juri Belikov and Alberto-Jesus Perea-Moreno

Received: 30 January 2026
Revised: 11 March 2026
Accepted: 19 March 2026
Published: 21 March 2026

1. Introduction

According to the International Energy Agency's report, *Energy Efficiency 2024*, the energy demand in the building sector was over 120 exajoules (EJ) in 2023, accounting for 28% of the global final energy consumption. It has increased by an average of 0.9% per year between 2010 and 2023 [1]. Buildings not only account for a large share of energy consumption but also have a large potential for energy savings. Appropriate energy-saving measures can reduce energy consumption in buildings by around 20% on average [2,3]. Therefore, improving building energy performance has become an important research focus.

In recent years, ML has been extensively studied through review articles in the field of building energy management [4], building energy prediction [5], fault detection and diagnosis [6], and, again, building energy management [7,8]. These recent studies provide a detailed summary of prior work while noting that the practical application of ML re-

mains limited. A likely reason lies in data quality, which can be improved through data preprocessing [5,9].

There is a growing consensus among ML researchers that data quality and preprocessing are not auxiliary steps but fundamental components of successful ML models [10]. The importance of data and feature quality in ML has also been repeatedly emphasised. As noted by Hastie et al. [11], the success of a model is often determined more by the data and the features than by the learning algorithm itself.

However, in the building energy field, advanced studies on data preprocessing remain limited compared with the extensive work on complex optimisation and ML algorithms [12], highlighting the importance of data preprocessing studies and related literature reviews.

Figure 1 illustrates the general ML workflow in the building energy field. The ML pipeline can be divided into four stages: data collecting, data preprocessing, model training and model deployment.

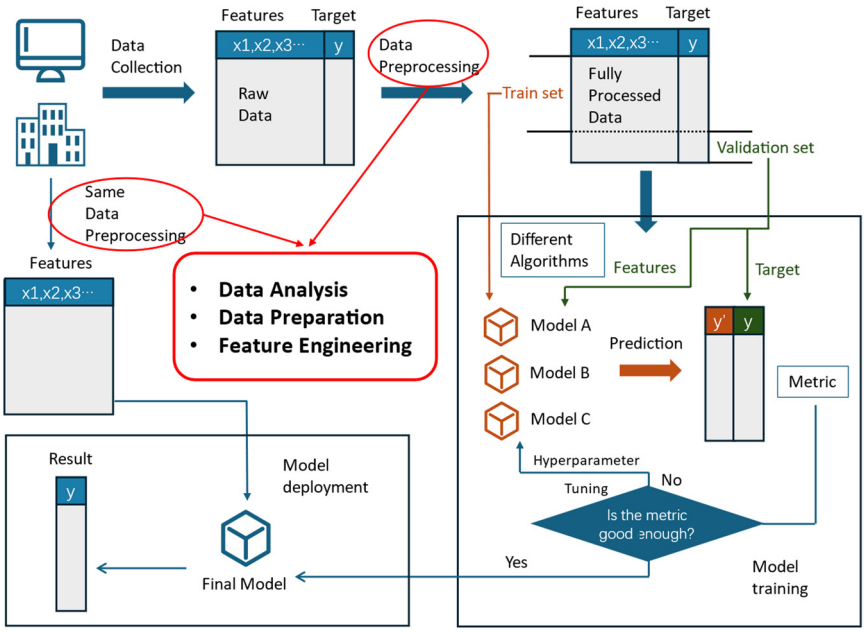

Figure 1. Pipeline of the general ML process in the building energy field.

Whether in the training stage or the usage stage of a model, data that are always known, such as x1 and x2, are called 'features'. The variables shown as 'y' in Figure 1 are only known during the model training stage and need to be predicted by the model in the model deployment stage; they are referred to as 'targets'.

As shown, data preprocessing serves as a critical bridge between real-world measurements and the ML algorithms. Features $x1, x2, \ldots, xn$ and target values y are collected through sensors in the building or typically labelled by engineers and transformed into a standardised dataset during data preprocessing.

If the target is a continuous variable, the task may involve time-series regression, such as forecasting hourly building electricity consumption, or general regression, such as predicting system parameters or heating load for specific conditions. In addition, when the target is discrete, the task becomes classification, for example, identifying fault types or the operating status of a heating, ventilation and air conditioning (HVAC) system.

Typically, in the building energy field, depending on the task and the specific form of the data, different preprocessing techniques are required. The data types can be categorised as follows:

- Time: Shown as the string 'YYYY-MM-DD' or 'HH: MM: SS', representing the collection time of the data. In the building energy field, parameters always have a relationship with time.
- Level: Shown as an integer '0, 1, 2', such as the fault level of the HVAC system. The numerical relationship has a corresponding physical meaning.
- Classification/Status: Shown as integer, string or character, such as open status 'ON/OFF' or Climate Area 'A, B, C, D, E'.
- General Data: The real numbers whose numerical precision is determined by sensors have actual physical meaning, and many have theoretical value ranges. For example, 'outdoor air temperature' should not be more than 50 °C at any time.
- Invalid Data: Illegal format data caused by sensor or network failure. Such as 'NaN' data, which stands for 'Not a number' or simply missing data.

In parallel with the growing attention to data preprocessing in recent years, several review studies within the building energy domain have begun to discuss preprocessing as part of the ML workflow.

As for the building load or energy consumption prediction research, Wang, Z. [13] reviewed the ML practical applications in the full cycle of studies in building energy, including the design, operation, maintenance, and retrofit, and discussed the data preprocessing steps as a future direction. In addition to that, Zhang, L. et al. [5] discussed ML applications in building load prediction from the sensor level to the data level, including data preparation and feature engineering.

Unlike previous reviews, Sun, Y. et al. [14] summarised the entire ML process in building energy consumption prediction, including data preprocessing and algorithmic applications. Also, Hou and Evins [15] summarised the workflow, including data preprocessing based on Neural Network ML algorithms. Furthermore, Zhang, Y. et al. [16] proposed a technical framework including data preprocessing that effectively integrated the advanced ML algorithms, language models, and building energy simulation models.

After producing numerous summaries and studies, researchers began to focus on the detailed processes of ML application in the building energy field. Liu, H. et al. [17] summarised the influencing parameters and time scales of ML and found that data preparation techniques can effectively solve time-scale problems that commonly arise in building energy prediction.

Based on existing review papers, Chen, Y. et al. [18] not only summarised the latest advanced algorithms, but also explained the algorithms, as well as feature engineering and clustering from a theoretical perspective.

Some researchers focused on feature engineering. Chen, G. et al. [19] not only explored an ML framework in building energy consumption prediction but also introduced four efficient feature extraction techniques. Focusing on occupancy identification and prediction, as well as window-opening detection, Dai, X. et al. [20] summarised the most suitable features and targets for different building operating conditions. Also, Wang, Z. et al. [21] presented a systematic review on feature engineering for building energy consumption prediction, outlining the current state and limitations of research.

In summary, research on ML in the building energy field has generally evolved through three stages: algorithm application, framework development, and data preprocessing. However, most algorithm-focused review articles treat preprocessing techniques merely as directions for future work, whereas framework-focused reviews consider them only as secondary or supporting components. For those review articles that focus on data preprocessing, the entire stage is missed, while only specific techniques are mentioned.

Therefore, in the existing literature reviews on using ML for improving building energy performance, there is a research gap in summarising all data preprocessing techniques,

which is fulfilled by the present study as a novel contribution in the field. Different from previous works, this review systematically covers all stages of data preprocessing across diverse ML applications for building performance.

The review is guided by three research questions:

- (RQ1)—What data preprocessing techniques have been applied in ML-based building energy performance improvement studies, and how can they be systematically categorised as a workflow?
- (RQ2)—What evidence exists on the comparative strengths and limitations of these techniques across different tasks and data types?
- (RQ3)—Where do gaps, misconceptions, or underexplored techniques remain in current practice?

To answer these questions, a literature search has been conducted. After screening and eligibility checks, seventy-three articles were included. These studies were coded and synthesised according to the presented workflow.

The substantive contributions can be summarised as follows:

- A unified data preprocessing workflow is presented. The workflow structures the data preprocessing stage into three steps, clarifying which operations are essential and which depend on the task or dataset characteristics.
- Techniques for all steps are systematically reviewed in comparison with evidence from the literature, highlighting where consensus exists and where results remain inconsistent.
- Corrective guidance on some data preprocessing techniques is provided in this article by comparing different works. In addition, some techniques that have not been thoroughly explored are identified.

The remainder of the paper is organised as follows. Section 2 details the methodology. Section 3 analyses the three layers of preprocessing: data analysis, data preparation, and feature engineering. Section 4 discusses key findings, and Section 5 concludes with recommendations for future studies.

2. Methodology

For this review, a systematic literature search was first conducted utilising the Scopus and Web of Science databases (2015–October 2025), chosen for their comprehensive coverage of peer-reviewed scientific publications across diverse disciplines. The query of titles, abstracts and keywords was designed as follows:

- “Machine Learning” OR “ML” OR “AI” OR “Artificial intelligence” OR “digital-twin”

  AND

- “Energy” OR “HVAC” OR “Air-conditioning” OR “Air condition” OR “Heating” OR “Cooling”

  AND

- “Building”

  AND

- "Preprocessing" OR “feature engineering”

Figure 2 illustrates the annual publication trend of journal and review articles retrieved from the databases based on the defined query. Relevant studies began to emerge in 2017. During 2016 and 2020, the number of publications remained relatively low, suggesting that research on data preprocessing in the building energy domain was still in its early developmental stage. A pronounced increase in publication output can be observed from 2021, indicating growing academic interest in this topic. Therefore, this review primarily

focuses on studies published from 2020 to October 2025, capturing the period of rapid development and intensified research activity.

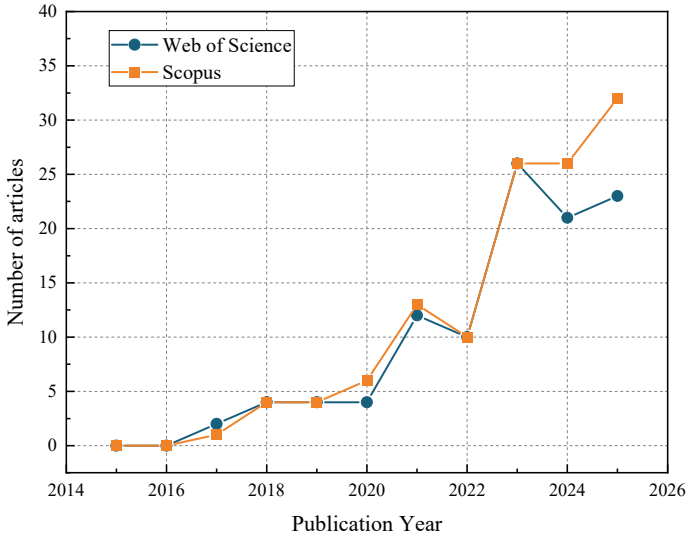

Figure 2. Annual publication trend of selected articles (2015–2025).

The articles identified were processed through the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology. The detailed selection procedure is shown in Figure 3. A total of 209 records were initially identified. After removing 81 duplicate records, 128 unique articles remained for screening. During the title and abstract screening stage, 30 records were excluded due to their irrelevance to the scope of this review. Subsequently, 98 full-text articles were accessed for eligibility. Of these, 30 articles were excluded for the following reasons: (1) lack of explicit description of data sources or data preprocessing techniques (n = 13); and (2) focus on urban-scale building energy systems or supply-side/grid-level research rather than building-level applications (n = 17). Ultimately, 73 studies (63 journal articles, 10 review articles) were included in this review.

The inclusion criteria were defined as follows: (1) studies focusing on building energy research, including thermal energy behaviours and electricity energy; (2) explicit description of data preprocessing techniques applied before ML modelling; (3) peer-reviewed journal articles and review articles published between 2020 and October 2025.

The exclusion criteria were: (1) studies about building energy but at urban scale or grid level; (2) studies lacking description of data preprocessing steps or the dataset; (3) conference papers, theses, and technical reports; (5) articles published outside the defined database or time range (including early access). Further information about the PRISMA checklist can be found in Supplementary Materials Table S1.

Table 1 organises the selected journal articles by dataset, building type, and ML task. It shows the diversity of data sources (measured, simulated, or statistical), a range of building contexts from residential to office and mixed-use, and tasks covering regression, classification, and clustering. The selected review articles are analysed in the introduction section [5,13–21]. The prediction targets vary from energy consumption to indoor thermal comfort and fault diagnosis. This overview sets the context of the review before the detailed analysis of distributions in the following sections. Since the objective of Table 1 is to summarise the application of preprocessing techniques in original research studies, review articles were not incorporated into the tabulated comparison. However, when certain preprocessing techniques were discussed in review articles but not explicitly reported in

the collected journal studies, they were additionally referenced and discussed in Section 3 to ensure conceptual completeness.

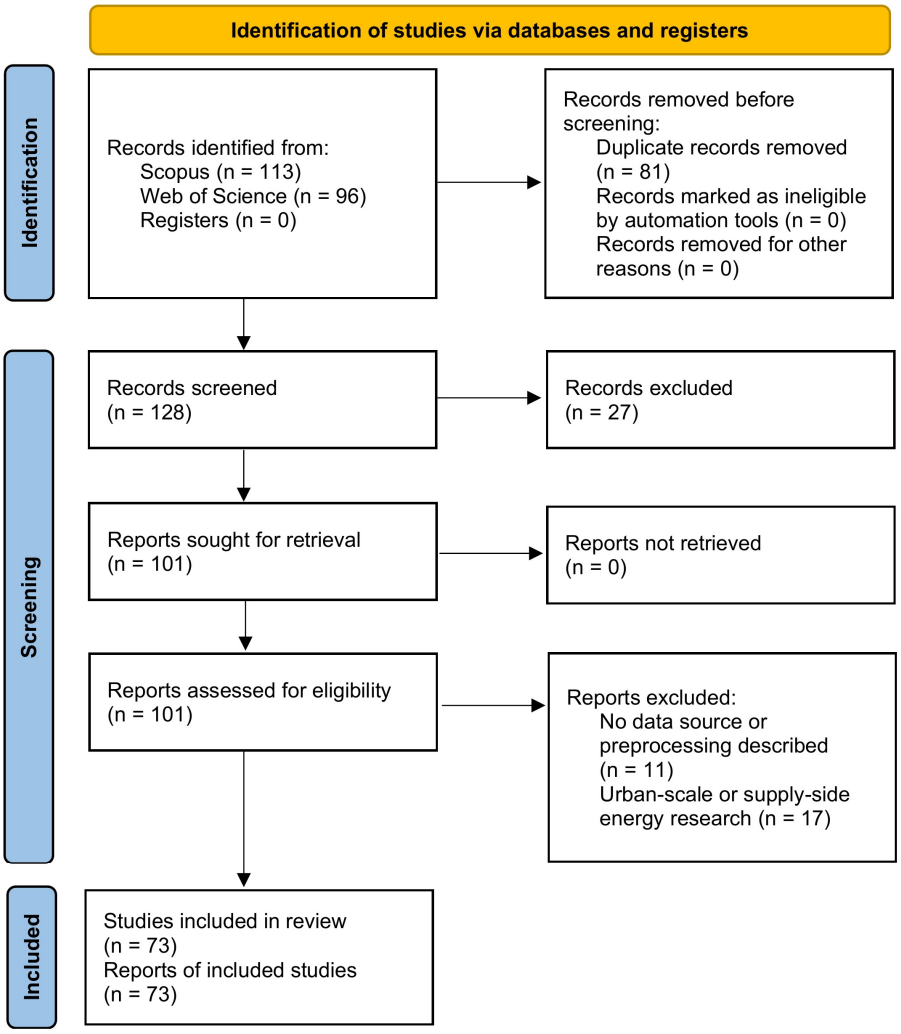

Figure 3. PRISMA process.

The distribution of studies by type of data, type of building, prediction target, and ML task is shown in Figure 4. Most studies are carried out on the measured data, which is most likely to be noisy. As for the building types, residential buildings are the most frequently studied, followed by office and mixed-use buildings. The numerical differences among the building types are not very significant. In terms of prediction targets, the preprocessing of electricity consumption is by far the dominant focus. Preprocessing discussions are more detailed in electricity consumption modelling than in heating or cooling load forecasting, as well as other topics. Lastly, for ML tasks, time-series regression was considered independently from regression in this review, because of the different requirements for data preprocessing. Fewer studies address classification, such as anomaly detection tasks.

Table 1. Scope of collected journal articles.

Ref.	Public Dataset	Data Type	Scale	Building Type	System	Prediction Target	Timestep	ML Task	Year
[22]	Yes	Measured Data	Building	Mix	-	PMV	-	Regression	2025
[23]	No	Measured Data	Building	Residential	-	Electricity Consumption	Hourly	Time-Series Regression	2022
[24]	No	Measured Data	Building	Commercial	-	U-Value	-	Clustering	2021
[25]	No	Measured Data	System	Office	PV System	PV Electricity Output	Hourly	Time-Series Regression	2023
[26]	Yes	Measured Data	Building	Commercial	-	Electricity Consumption for Heating/Cooling	Annual	Regression	2022
[27]	Yes	Simulation	Building	Mix	-	Heating & Cooling Load per Area	-	Regression	2025
[28]	Yes	Simulation	Building	Mix	-	Heating & Cooling Load per Area	-	Regression	2024
[29]	Yes	Measured Data	Building	Mix	-	PMV	-	Regression	2024
[30]	Yes	Statistical Data	Building	Commercial	-	Electricity Consumption	Annual, Weekly, Hourly	Regression	2025
[31]	No	Measured Data	System	Residential	Variable Refrigerant Flow System	Refrigerant Leak Level	-	Classification	2024
[32]	1. No 2. Yes	1. Measured Data 2. Measured Data	1. Building 2. System	1. Dormitory 2. Residential	1. - 2. Appliances	Electricity Consumption	Hourly	Time-Series Regression	2024
[33]	1. Yes 2. No	1. Measured Data 2. Measured Data	1. Building 2. Building	1. Office 2. Office	1. - 2. -	Electricity Consumption	Hourly	Time-Series Regression	2021
[34]	No	Measured Data	System	Residential	Appliances	Average Electrical Power	5 min	Time-Series Regression	2022
[35]	Yes	Simulation	Building	Mix	-	Heating & Cooling Load per Area	-	Regression	2020
[36]	No	Measured Data	Building	Commercial	-	Electricity Consumption	Hourly	Time-Series Regression	2025
[37]	1. No 2. Yes	1. Simulation 2. Measured Data	1. System 2. System	1. Office 2. Office	1. HVAC 2. HVAC	Electricity Consumption	Hourly	Time-Series Regression	2023
[38]	Yes	Simulation	System	Office	HVAC	Fault Level	-	Classification	
[39]	Yes	Measured Data	Building	Mix	-	Primary Use Electricity	Hourly	Time-Series Regression	2022
[40]	Yes	Statistical Data	Building	Residential	-	Energy Efficient Level	-	Classification	2023
[41]	Yes	Measured Data	System	Residential	Appliances	Electricity Consumption	Hourly	Time-Series Regression	2023
[42]	Yes	1. Measured Data 2. Measured Data	1. Building 2. Building	1. Residential 2. Commercial	1. - 2. -	Electricity Consumption	Hourly	Time-Series Regression	2024
[43]	No	Measured Data	System	Apartment	Window	Window Open Status	-	Classification	2023
[44]	No	1. Measured Data 2. Measured Data 3. Measured Data	1. Building 2. Building 3. Building	1. Educational 2. Educational 3. Dormitory	1. - 2. - 3. -	Electricity Consumption	Hourly	Time-Series Regression	2023
[45]	Yes	Measured Data	System	Residential	Appliances	Electricity Consumption	Minutely, Hourly, Daily, Weekly	Time-Series Regression	2021

Table 1. *Cont.*

Ref.	Public Dataset	Data Type	Scale	Building Type	System	Prediction Target	Timestep	ML Task	Year
[46]	No	Measured Data	System	Office	HVAC	Cooling Load	Hourly	Time-Series Regression	2024
[47]	Yes	Statistical Data	Building	Residential	-	Electricity Consumption	Annual	Regression	2021
[48]	No	Measured Data	Building	Commercial	-	Cooling Load	Hourly	Time-Series Regression	2024
[49]	Yes	Simulation	Building	Apartment	-	Electricity Consumption	Hourly	Time-Series Regression	2025
[50]	1. Yes 2. Yes 3. Yes	1. Measured Data 2. Measured Data 3. Measured Data	1. Building 2. Building 3. Building	1. Educational 2. Office 3. Residential	1. - 2. - 3. -	Occupancy State	-	Classification	2023
[51]	No	Measured Data	Building	Livestock	-	Indoor Temperature	Hourly	Time-Series Regression	2024
[52]	Yes	Statistical Data	Building	Mix	-	Thermal Comfort Level	-	Classification	2024
[53]	No	Measured Data	System	Residential	Window	Window Open Status	-	Classification	2024
[54]	No	Measured Data	System	Industrial	HVAC	Fault Level	-	Classification	2023
[55]	No	Measured Data	Building	Educational	-	Electricity Consumption	15 min	Time-Series Regression	2021
[56]	Yes	Simulation	System	Office	HVAC	Electricity Consumption	Hourly	Time-Series Regression	2023
[57]	Yes	Simulation	Building	Mix	-	LEED credits	-	Regression	2025
[58]	No	Measured Data	System	Cold Storage	Air-conditioner	Electricity Consumption	Hourly	Time-Series Regression	2024
[59]	No	Measured Data	Building	Hospital	-	Electricity Consumption	15 min	Time-Series Regression	2023
[60]	Yes	Simulation	Building	Residential	-	Heat Flux	-	Regression	2023
[61]	No	Measured Data	Building	Hospital	-	Electricity Consumption	Hourly	Time-Series Regression	
[62]	1. Yes 2. Yes	1. Measured Data 2. Measured Data	1. System 2. System	1. Residential 2. Residential & Industrial	1. Appliances 2. Appliances	Electricity Consumption	0.02 s	Time-Series Regression	2023
[63]	1. No 2. Yes	1. Measured Data 2. Measured Data	1. Building 2. Building	1. Educational 2. Mix	1. - 2. -	Electricity Consumption	Hourly	Time-Series Regression	2025
[64]	Yes	Measured Data	Building	Residential	-	Electricity Consumption	Hourly	Time-Series Regression	2025
[65]	No	Statistical Data	Building	Residential	-	Electricity Consumption	Hourly	Time-Series Regression	2024
[66]	Yes	Measured Data	Building	Mix	-	Electricity Consumption	Hourly	Time-Series Regression	2023
[67]	Yes	Measured Data	Building	Office	-	Electricity Consumption	Hourly	Time-Series Regression	2022
[68]	No	Measured Data	Building	Mosque	-	Electricity Consumption	Hourly	Time-Series Regression	2023
[69]	No	Simulation	Building	Office	-	Electricity Consumption per Area, Carbon Emission per Area, Cost per Area	-	Regression	2023
[70]	Yes	Simulation	Building	Residential	-	Heating & Cooling Load per Area	-	Regression	2020
[71]	Yes	Measured Data	Building	Residential	-	Electricity Consumption	Minutely, Hourly, Daily, Weekly	Time-Series Regression	2025
[72]	Yes	Measured Data	System	Residential	Appliances	Electricity Consumption	30 min, Hourly	Time-Series Regression	2023

Table 1. *Cont.*

Ref.	Public Dataset	Data Type	Scale	Building Type	System	Prediction Target	Timestep	ML Task	Year
[73]	Yes	Simulation	Building	Residential	HVAC	High/Low Energy Consumption	-	Classification	2025
[74]	No	Measured Data	Building	Mix	-	Supplied Heat Flow	Hourly	Regression	2025
[75]	No	Measured Data	System	Residential	Appliances	Electricity Consumption	Hourly	Regression	2025
[76]	1. No 2. No 3. No	1. Simulation 2. Measured Data 3. Measured Data	1. Building 2. Building 3. Building	1. Office 2. Mix 3. Residential	-	Electricity Consumption	Hourly	Time-Series Regression	2025
[77]	No	Simulation	Building	Educational	-	Indoor Temperature	Minutely	Time-Series Regression	2023
[78]	No	Measured Data	Building	Residential	-	Indoor Temperature	Daily	Regression	2025
[79]	No	Simulation	Building	Residential	-	Energy Demand, PV Electricity Generation	Hourly	Time-Series Regression	2024
[80]	1. No 2. No	1. Measured Data 2. Measured Data	1. Building 2. Building	1. Residential 2. Residential	-	Energy Demand, Energy Generation	Hourly	Time-Series Regression	2025
[81]	Yes	Simulation	Building	Residential	-	Heating & Cooling Load per Area	-	Regression	2024
[82]	No	Measured Data	Building	Mix	-	Energy Consumption Index of Insulator	-	Regression	2025
[83]	No	Measured Data	Building	Mix	-	Electricity Consumption	15 min	Regression	2025
[84]	Yes	Measured Data	Building	Mix	-	Room Inhabitancies	30 s	Classification	2025

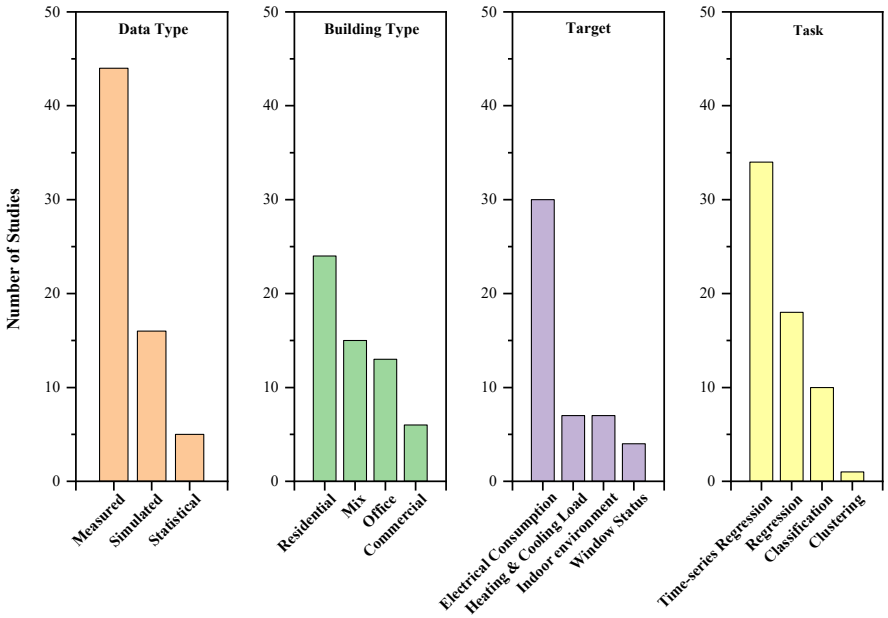

Figure 4. Distribution of studies in the review.

To characterise the preprocessing techniques adopted in the selected studies, a workflow-based framework is proposed, categorising the techniques according to their functional position within the ML pipeline. The data preprocessing can be organised based on 3 main steps:

- Data Analysis: Exploratory Data Analysis (EDA) serves as the core method for understanding dataset characteristics, supported by two additional analytical methods—clustering (either time-based or algorithm-based) for pattern recognition, and temporal statistical tests for checking data stationarity.
- Data Preparation: Includes outlier detection, missing value imputation, data transformation and data sampling. Data preparation aims to include both features and targets.
- Feature Engineering: Includes feature extraction and feature selection. Feature extraction creates new features, while feature selection only selects features from the existing feature space.

Figure 5 illustrates the proposed data preprocessing workflow. Each step consists of several methods, which are further divided into specific techniques. Solid arrows indicate essential processes in the ML pipeline, while dashed arrows represent optional transitions.

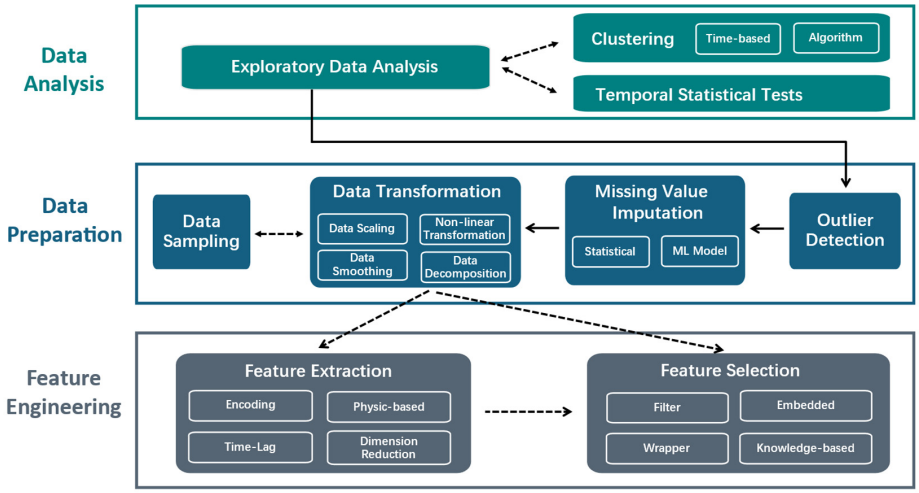

Figure 5. Workflow of preprocessing techniques for building energy data modelling.

In the following section, details describing the techniques are reported, mentioning the related studies adopting them.

3. Preprocessing Techniques

3.1. Data Analysis

This section introduces data analysis, which is an underemphasised component of the ML pipeline. The data from real buildings are often complex, incomplete, and vary over time, requiring an EDA for a better scope. The hidden patterns under different building working conditions can be uncovered by the clustering algorithm. In building energy time-series regression tasks, such as energy or load prediction, temporal statistical tests are essential to assess data stationarity. Stationary data typically leads to more-reliable predictions. If the data is non-stationary, the corresponding data preparation should be applied.

3.1.1. Exploratory Data Analysis

EDA is the work of inferring information from the dataset to enable the ML model to work in the most efficient way possible [85].

As shown in Figure 6, EDA has three progressive stages: data profiling, data visualisation for structural insights, and statistical relationship analysis [86]. Data profiling examines variable types and missing value proportions. Data visualisation reveals distribution patterns, outliers, and potential non-linear trends. Statistical relationship analysis quantifies correlations among variables. The insights derived from EDA serve as diagnostic signals that determine whether subsequent preprocessing steps are required.

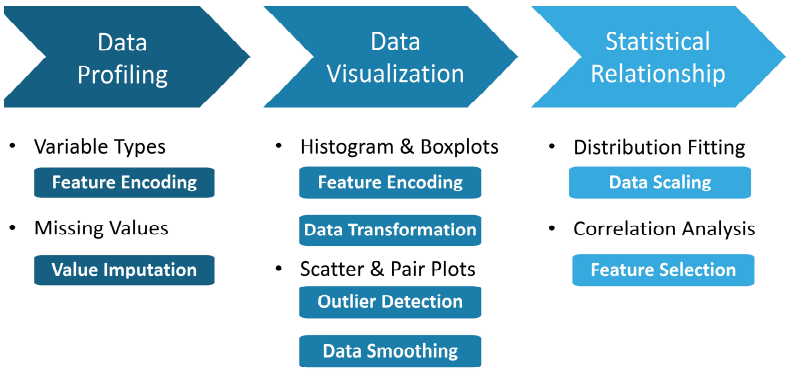

Figure 6. Framework of Exploratory Data Analysis.

It is found that the researchers' EDA can be broadly categorised into three levels. The first level is the most basic textual introduction, sometimes with a basic visualisation of the data. The second level has a more detailed statistical analysis, as well as other visualisations. The third level follows the complete logic and flow of EDA the most.

For the first level, Kumar Mohapatra, S. et al. [26] broadly described the building energy data, the number and type of buildings, and linear correlation relationships among features like heating degree days and cooling degree days. Klemp, S. et al. [24] mentioned the data source with a simple text and showed the distribution histogram of window-to-wall ratio data.

For the second level, Jeong, J. et al. [25] summarised the information, listing the PV generator time intervals, value domains, and units of weather data collection, and analysed different weather conditions. Ullah, F.U.M. et al. [45] plotted lines and distribution histograms on different time scales for the target value, as well as completed group-size analysis for all housing appliances' features. Alrashidi, M. et al. [68] performed a time-

series comparison of cooling loads and temperatures, found the time delay, and plotted correlation heat maps.

As for the third level, Mahmood, S. et al. [28,52] completed an EDA with multiple plots and heat maps of feature correlations in both thermal comfort and heating/cooling load, which provided reasons and theoretical support for all the later data preparation and feature engineering. Rahmanparast, A. et al. [22] analysed the missing rate, mean value, and range of values. The distribution of the variables was plotted, as well as the correlation heat map of features and the thermal comfort index, Predicted Mean Vote (PMV). In the data preparation stage, they chose the techniques according to the results of EDA.

It was found that less than 20% of researchers introduced 'EDA' as a technique. This indicates that researchers should place greater emphasis on a systematic data analysis process.

3.1.2. Clustering

Clustering analysis has already been applied in building energy studies to identify energy performance patterns. It can assist researchers in classifying climate zones and developing region-specific energy retrofitting strategies for residential buildings [87]. In ML-based building energy applications, clustering can be used as a data preprocessing technique to segment data into groups with similar operational characteristics. For example, clustering can distinguish operational modes (e.g., peak vs. off-peak hours), allowing for the development of specialised ML models for different modes, thereby improving accuracy [88,89].

According to the studies collected, clustering algorithms can be classified as follows [90]:

- Hierarchical: Agglomerative
- Partition-based: K-means
- Density-based: Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Mean Shift, and Shared Nearest Neighbours (SNN)
- Probabilistic: Fuzzy C-means
- Customised: Validation-loss-based dynamic clustering

A comparison of K-means, Agglomerative, DBSCAN, Mean Shift and SNN processing of data from a commercial building has been made by Klemp, S. et al. [24]. They found that after Box–Cox data transformation, DBSCAN performs best on metrics, but categorises many data samples as noise. The overall combined performance is best for the SNN algorithm. Among all the algorithms, Mean Shift performed the worst. Weber, S.A. et al. [74] used K-means clustering to separate the district heating fluid volume flow data to distinguish the operational modes.

For the ML algorithm, Bayesian linear regression and its ensemble model, Fuzzy C-means improved the accuracy of predicting the building energy consumption [33].

Jeevakarunya and Manikandan [30] applied K-means to segment time data into broader categories, such as daily periods, weekdays versus weekends, and seasonal cycles. While Liu, Y. et al. [46] made a time-based operational-mode classification of building cooling load. Using time data information in the dataset, the load curve could be easily categorised into weekdays, weekends and holidays.

Unlike all other researchers, Tang, L. et al. [66] proposed using validation-loss-based dynamic clustering. While there is a data confidentiality request on the building energy consumption prediction model deployment stage, this technique can implement the cooperation among multiple buildings with insufficient data while preserving privacy.

It is shown that clustering studies are always carried out in office or commercial buildings [76], which have a more significant operation pattern than residential buildings. Algorithms such as K-means or DBSCAN offer greater flexibility and can automatically

identify patterns based on actual data behaviour. However, they require careful hyperparameter tuning and may result in clusters that are harder to interpret. In contrast, time-based clustering [83] is simple to use and explain with the schedules of occupancy or HVAC systems in buildings. Yet it may fail during holidays or unexpected events, leading to a wrong result [91].

3.1.3. Temporal Statistical Test

For time-series regression tasks, like cooling and heating loads forecasting or building energy consumption prediction, a temporal statistical test is mandatory. Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Augmented Dickey–Fuller (ADF) tests are two classical algorithms to test the stationarity of time-series and are commonly used before ML modelling [92].

If the test results show that the sequence is stationary, then it can be modelled directly using ML algorithms such as Auto Regressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), etc. If the time-series results are realistically non-stationary, then processing such as logarithmic transformation or differencing needs to be performed on that data [93].

A few articles about temporal statistical tests are collected, even though there are more than 30 time-series regression tasks.

The combination of both KPSS and ADF was used in the study by Chou et al. [94] to classify the time-series into stationary, trend-stationary, difference-stationary, or non-stationary categories. Based on the classification, appropriate preprocessing strategies were determined, including detrending, differencing, or applying decomposition techniques.

3.2. Data Preparation

This section presents a series of data preparation techniques that modify the dataset without changing the number and meaning of features. Effective data preparation influences not only ML model accuracy during training but also the robustness and reliability of the model in practical building energy improvement. Improper data preparation is a key reason for low accuracy and for discrepancies between training and real-world results [12].

3.2.1. Outlier Detection

Outliers may interfere with later processing steps, so they should be detected and handled first. Under some conditions, the researchers only deleted the rows where a missing value existed, but forgot to detect the outliers [57].

A substantial number of data samples may be removed in certain building energy datasets [65], highlighting the importance of outlier detection. The interquartile range (IQR) with a boxplot is always used for outlier detection on features like indoor/outdoor temperature [31], and also for the whole dataset [28,83].

Besides IQR, other algorithms have been used for building energy consumption prediction. Qiao, Q. et al. employed Local Outlier Factor (LOF) [44], which determines the outliers by the density of their neighbouring data. The generalised extreme studentized deviation (GESD) is also used by researchers such as [49]. GESD is statistically based, and it is particularly suitable for time-series data in a normal distribution.

After considering the collected articles, the algorithms for detecting outliers are as follows:

- IQR: The most popular and useful one, which is often accomplished with the help of boxplots that can reflect the data dispersion and show outliers.
- LOF: Suitable for high-dimensional data (multiple sensors) and not a time-series regression task.

- GESD: Suitable for small samples with approximately normally distributed data with a time-series regression task.
- Rule-based outlier detection: Setting the acceptable data range by expert knowledge.

A specific rule-based outlier detection can be used on the building setpoint temperature [56] or indoor/outdoor temperature [51]. For general parameters, the default baseline value and its maximum permissible percentage error are set as the rules for outlier detection [56].

In certain cases, retaining outliers is justified when they have physical significance. Asadi, N. and Moosavi, L. [53] used IQR and found that the identified outliers represented the typical extreme weather conditions. Considering climate change, they chose to keep them to prevent the ML model failure when encountering extreme weather again.

IQR was the most frequently adopted outlier detection method in the reviewed studies. More comparative analysis of the detection algorithms used on building energy data needs to be undertaken.

3.2.2. Missing Value Imputation

When performing time-series regression tasks, it always requires the data to be continuous. Due to data corruption and missing and removed of outliers, appropriate imputation techniques are required to complete the dataset [95]. The techniques can be mainly classified as statistical and ML model-based imputation.

Statistical imputation refers to the techniques that fill the missing values using fixed values, local data, or an interpolation fitting function. ML model-based imputation refers to the techniques that use ML algorithms to estimate the missing value.

For the temporal data, such as building energy consumption or weather conditions, the forward/backwards filling technique can be sensible. The technique is to fill the missing blank with the previous/next value [45,48,71]. A similar way to fill the blank with local values is interpolating the missing value as the average of the surrounding values [31,49,74,77]. Furthermore, Pai, H.A. et al. [75] used variable-specific imputation on a household energy consumption and weather dataset.

Using the Artificial Neural Network (ANN) and LSTM algorithms, Buțurache and Stancu [39] compared the linear and the cubic spline imputation, which uses a linear or cubic function fitting locally beside the missing data. The results show that the cubic spline imputation worked better.

Hussain, A. [36] exhaustively summarised a wide range of imputations using LSTM in building electrical load forecasting under different missing value percentages. For statistical imputation, using mean imputation, median imputation and Multiple Imputation by Chained Equations (MICE) performs better than zero imputation and mean imputation alone. For ML-based imputation, Gradient Boost is the best one among several algorithms. And as a hybrid imputation, the combination of Gradient Boost and MICE performs a lot better than any other algorithm. In addition, the missing rate should be detected to select the appropriate imputation technique. For a missing rate that is less than 30%, the statistical imputation techniques can be considered; otherwise, ML model-based imputation performs better.

Different imputation techniques have different proper uses [96]. Fixed-value imputation can be used for data with less disturbance and missing values, such as indoor temperature. Data-based imputation can be used for features with strong time continuity, such as the open status of HVAC systems. As for interpolation, it is more suitable for weather information that has periodic or smooth physical signals. Different ML imputations perform differently according to the dataset, while simple models like k-Nearest

Neighbours (KNN) perform well under a medium-missing-ratio dataset [83], and complex models [78] perform better under a high-missing-ratio dataset [97].

3.2.3. Data Transformation

Data transformation aims to reshape and scale the building dataset to make it suitable for the ML algorithm. These transformations have various purposes: adjusting the scale and distribution of features, smoothing fluctuations and removing noise. Data decomposition is typically applied to the target rather than the features to decompose complex temporal patterns into simple components.

Data Scaling

In building operation data, data often have different scales: temperature changes may be within 20 °C, but the collected power data may jump by thousands of Watts. To eliminate the impact of inconsistent data scales on the results, normalisation or standardisation is used to preprocess the data. Scaling significantly affects the performance, and the wrong scaling can be even worse than not scaling the data at all [98]. For the collected works, around 50% of the articles did not mention the data scaling or normalisation.

Min-max normalisation has been widely used [24,27,28,30,35,37,50,60,74,79,80,84] in the building energy field. It can map the value to [−1, 1] or [0, 1]. Min-max normalisation is highly sensitive to extreme values; therefore, the outlier detection must be processed before using min-max normalisation. Gao, Y. et al. [99] used Z-score standardisation on the office building energy data, and Almadhor, A. et al. [73] used the same standardisation on residential building data.

Yang and Sung [62] compared different data scaling techniques, such as min-max, MaxAbs, Z-score, L1 regularisation and Robust Scaler together in detail while using several types of ML algorithms on residential and industrial buildings. In most cases, min-max normalisation showed the best performance.

Non-Linear Transformation

Non-linear transformation, such as the Box–Cox and Yeo–Johnson transformation, is often used to improve data distribution skewness issues.

While analysing the building geometry feature window-to-wall ratio, Klemp, S. et al. [24] applied the Box–Cox transformation to address the right-skewed patterns in the dataset.

Yang and Sung [62] used the Yeo–Johnson transformation on appliance data of residential and industrial buildings. The results show that it is not as effective as the min-max transformation, but it is the second-best choice among all other cases.

The non-linear transformation techniques and the detection of data skewness remain underutilised, and further research is needed to clarify their potential benefits.

Data Smoothing

Due to sensor errors and disturbances existing in buildings, sometimes it is necessary to handle the noise and disturbances in the target variables (energy consumption and heating/cooling loads) with data smoothing techniques. Unlike imputation, the smoothing process changes existing data that exhibit abnormal fluctuations.

Simple moving average (SMA) and exponential moving average (EMA) could be used as classic techniques. The principle is to shift a sliding window across the data and perform calculations on the values in the window to obtain the smoothed value. In SMA, all data points have equal weights, resulting in slower responsiveness, while EMA, as an improvement of SMA, introduces a smoothing factor to give higher weight to the data nearby, which is more suitable for the building cases.

The Savitzky–Golay (S-G) filter is an improvement on SMA and EMA based on local polynomial least squares fitting weight in the sliding window. It can filter out noise while ensuring that the shape and width of the signal remain unchanged [100].

The Kalman filter assumes that the errors in the data follow a Gaussian distribution, and it can minimise uncertainty in the data, which is particularly suitable for preprocessing building energy consumption data [101].

Due to the low accuracy of individual activity record data, SMA is used to smooth the record [71,72]. For building energy consumption data, SMA, S-G filter and Kalman filter methods are used [42,102].

Similarly, for fault detection and diagnosis of variable refrigerant flow systems, Essakali, N. et al. [31] compared SMA, S-G filters, and Kalman filters for temperature. Under four different ML algorithms, Kalman filters performed well, while SMA performed poorly in most cases.

Data smoothing techniques are summarised in Table 2. Several studies reported superior performance of the Kalman filter compared with other smoothing techniques. However, comparisons over different datasets remain insufficient.

Table 2. Summary of smoothing algorithms.

Principle	Algorithm	Drawback
Sliding windows	SMA	Slow responsiveness, high lags
	EMA	Easily affected by noises
	S-G	Risk of distortion
Minimising errors	Kalman filter	High calculation cost, complex

Data Decomposition

Data decomposition not only improves the practical application accuracy of the ML model but also enables engineers to better understand the trends and cyclical changes in the parameters during building operation.

Seasonal-Trend decomposition using Loess (STL) separates the input time-series into three additive components: trend, seasonal, and residual [103]. Therefore, it is specifically effective for identifying the periodic behaviour of a building. A suitable algorithm for non-stationary time-series is the wavelet transform (WT), which decomposes signals at different time and frequency scales.

Empirical Mode Decomposition (EMD) decomposes the noisy signals into a set of better-quality intrinsic mode functions. However, there could be a problem called 'modal mixing', where one function may contain multiple frequency components. To solve this problem, researchers proposed Ensemble Empirical Mode Decomposition (EEMD) [104].

Chou, S.-Y. et al. [93] made a comparison between EMD and WT while using the LSTM algorithm to predict the energy consumption of different buildings. The results showed that the model after WT was worse than the one without processing, while EMD improved the performance.

EMD can be used on different building consumption datasets to find the hidden patterns of energy consumption and select the related features [44]. In addition, Ruan, Y. et al. [23] compared the EEMD and Variational Mode Decomposition (VMD), and found that EEMD only partially solves the modal-mixing problem. Also, EEMD performed worse than VMD [105] under the ML algorithm Back Propagation Neural Network.

Jin, N. et al. [106] compared using Singular Spectrum Analysis (SSA), WT, EMD and VMD. They found that SSA performs best on an ML algorithm based on LSTM. Models using WT and EMD fluctuate significantly in areas expected to be stable and exhibit lag

during prediction. Models using VMD are sensitive to data changes and impose excessive penalties on boundary and internal jumps, which affects prediction accuracy.

Data decomposition can also be applied to features rather than the prediction target. De Rautlin de la Roy et al. [43] used STL on indoor temperature, humidity and carbon dioxide concentration.

Table 3 shows the summary of decomposition algorithms in the reviewed studies. There is no 100%-certain conclusion about the best decomposition technique. STL has the highest explainability. In some cases, WT-based models performed worse than models without decomposition. When combined with LSTM, SSA showed more stable performance compared with other decomposition techniques in the reviewed cases. While using EMD and EEMD, the problem of modal mixing should be tested.

Table 3. Summary of decomposition algorithms.

Algorithm	Frequency -Domain	Time -Domain	Prior Knowledge	Explainability	Trend Analysis
VMD	○	-	-	Low	-
EMD	-	○	-	Low	-
EEMD	-	○	-	Low	-
STL	-	○	○	High	○
SSA	-	○	-	High	○
WT	○	-	○	Mid	-

Note: '○' denotes the presence of the corresponding property, while '-' indicates absence.

3.2.4. Data Sampling

In building energy systems, data are often highly imbalanced: most of the records reflect normal operation, while faults are rare. For instance, an HVAC system may run correctly for over 95% of the time, with only occasional faults. If this imbalance is not addressed, an ML model can appear accurate simply by always predicting "normal", while in fact failing to detect any fault events. Data sampling is therefore essential to ensure the model learns to recognise these unusual yet important conditions.

Latin Hypercube Sampling (LHS) is a statistical sampling algorithm used to efficiently cover multidimensional parameter spaces [107]. While dealing with the task of generating a dataset for training an ML model from building energy simulation software, Saad, M.M. et al. [69] used LHS to generate the dataset covering diverse operating conditions. The LHS algorithm was also selected to expand the limited building energy simulation data by incorporating parametric uncertainty [108,109].

The Synthetic Minority Oversampling Technique (SMOTE) creates minority samples through interpolating between a data point and its nearest minority neighbours. SMOTE is used by Wang, Y. et al. [110] to solve the data imbalance with a few faulty data points of the air conditioning system. They found that using SMOTE may be more time-consuming and prone to overfitting.

Data imbalance was frequently observed but inconsistently addressed across studies.

3.3. Feature Engineering

Feature engineering can be categorised into feature extraction and feature selection. Feature extraction constructs new features, while feature selection only picks features from the original features. Feature engineering is not mandatory for any energy improvement task of buildings. But the high-dimensional and noisy building operational data make it a practically important step.

3.3.1. Feature Extraction

Feature extraction serves two primary purposes. The first is to transform incompatible data formats, such as label parameters of the system or time information collected by sensors, into a numerical format that can be the input of ML algorithms. The second is to extract the latent relationships or temporal dynamics among features, thereby improving the model's ability to learn and generalise.

Feature Encoding

In building energy applications, encoding is commonly applied to time-related variables (e.g., hour, weekday, month) and discrete states (e.g., on/off, fault level), to convert them into numerical representations suitable for ML task.

Considering the reviewed studies, three encoding techniques are observed, as Figure 7 shows: one-hot encoding, label encoding, and sine/cosine encoding. One-hot encoding expands each category into a binary feature. Label encoding assigns integer values to categories. And sine/cosine encoding is to transform the time into two periodic features, as shown in Equations (1) and (2).

$$T_{sin} = \sin\left(\frac{2\pi}{n} T\right) \quad (1)$$

$$T_{cos} = \cos\left(\frac{2\pi}{n} T\right) \quad (2)$$

where T represents the time and n = 24, 7, 12 for hour, weekday, and month, respectively.

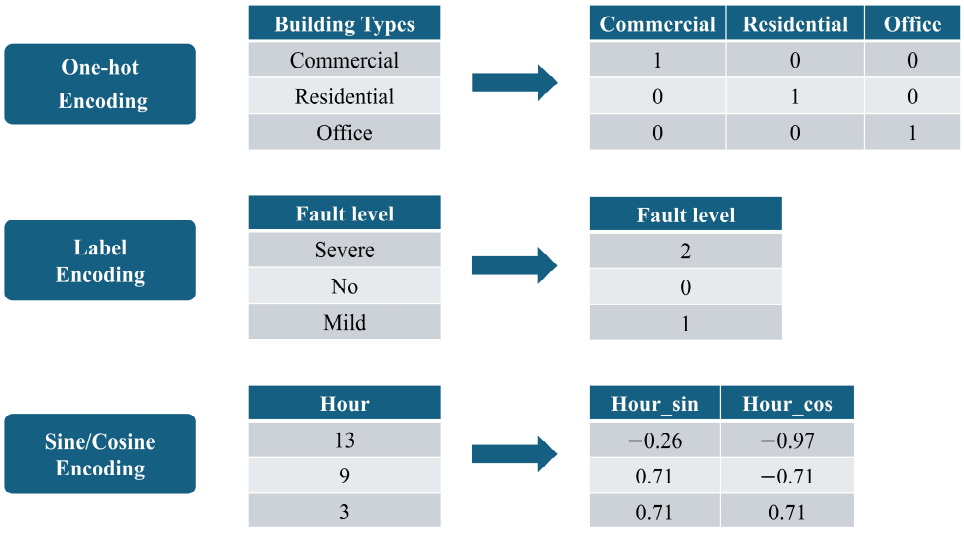

Figure 7. Encoding techniques.

In regression-based time-series forecasting, sine–cosine encoding is frequently used for hourly or weekly variables. Qiao et al. [44] reported improved correlation consistency after applying sine/cosine encoding. Alkhulaifi et al. [58] and Wang et al. [72] similarly adopted sine–cosine transformation to preserve temporal continuity.

By contrast, several studies applied label encoding to time variables without periodic transformation [41,45,51]. While computationally simple, label encoding may introduce artificial ordinal relationships (e.g., 23 and 0 treated as distant values).

For regression tasks with strong periodic patterns (e.g., hourly load prediction), preserving circular structure is commonly observed in recent studies [44,58,72]. For tree-based algorithms, label encoding is generally acceptable because model splits are threshold-based rather than distance-based [11].

In some cases (e.g., the level of building damage or fault level of a system), label encoding based on expert knowledge is most appropriate [111]. According to a recent FDD study of CO_2 heat pump systems, the one-hot encoded target representation can be transformed into a softened probability distribution during training to mitigate overfitting and enhance model generalisation [112]. For features with only a few classes, such as the season of the year, Perez Garcia, G.A. et al. [51] used one-hot encoding. Mansouri, A. et al. [57] introduced and discussed several encoding methods for predicting indoor temperature, including leave-one-out encoding and target encoding.

Time-Lag Features

For heating or cooling load prediction or energy consumption prediction tasks, features such as weather or human activity have delayed effects. Therefore, the time-lag features (features from some time earlier) should be extracted.

Time-lag features are of little meaning to time-series ML algorithms, such as Recurrent Neural Networks (RNNs) and LSTM, but these algorithms also require the length of the time sliding window, which is also a kind of time-lag. Therefore, determining the length of time-lag has become the biggest obstacle for researchers in dealing with time-series tasks.

From the articles that have been reviewed, it is shown that there are two main ways to assess the time-lag features: experience-based or analysis-based [113]. In the study of Alkhulaifi, N. et al. [58], the linear correlation of building cooling load and weather conditions over multiple preceding hours (and days) was calculated. Several time-lags with the highest correlation values were selected and introduced as new time-lag features.

Autocorrelation Analysis is used by Durand, D. et al. [34] to make a reasonable judgement. By plotting the autocorrelation function, they identified statistical correlation at specific lag intervals, which guide the selection of the time parameters.

Kim, J. et al. [63] directly plotted curves of different sine/cosine encoding time data together with the target, allowing them to intuitively obtain the appropriate time-lag setting.

Papias et al. [80] introduced not only multiple time-lag features ranging from 1 to 5 h, but also statistical values, including the Z-score and the 3 h rolling average.

Physics-Based Features

Unlike traditional computer science, which works on language or pixels, building data often has clear physical meanings. Since ML models may not capture these relations well, expert knowledge is useful for creating physics-based features that support ML models. Physics-based feature extraction can yield many variations depending on the expert knowledge and available data.

For building energy system features, Liu, Y. et al. [46] concluded that outdoor air temperature and cooling load are positively correlated, but the correlation changes significantly before and after the chillers start up. Therefore, they designed new features and achieved a less than 10% coefficient of variation for long-term predictions. Wang, H. et al. [72] specifically designed two features to account for the uncertainty in resident activities within the time-recording interval when modelling building energy demand.

For the weather features, Moon, J. et al. [32] utilised other researchers' studies ([114,115]) on the relationship between external weather conditions and heating and cooling load to calculate the discomfort index. Also, Jeong, J. et al. [25] proposed the new sky condition feature, which can represent the sky conditions at the site.

Based on the operational characteristics of building energy demand and on-site generation, Papias et al. [80] employed the cross-feature dependencies, such as flexibility above and below values. Such features embed domain knowledge into the ML process and enhance the performance of the predictions.

For general features, Sajjad, M. et al. [70] paired the features two by two, and then constructed combinations of up to sixth order, such as $\{x, y, x^2, xy, y^2, x^3, \ldots, xy^5, y^6\}$.

Dimension Reduction

Dimension reduction aims to decrease the number of input features while preserving the most informative ones. It is typically applied when high-dimensional inputs increase model complexity or risk overfitting [77]. The two most commonly used techniques in building energy studies are Principal Component Analysis (PCA) and AutoEncoders.

PCA performs a linear transformation of the original variables into orthogonal principal components that maximise variance along successive directions. Kumar Mohapatr, S. [26] used PCA for dimensionality reduction without reporting the resulting feature size. Durand, D. [34] used PCA on the selected features from multiple appliances and plotted the explained variance ratios for each principal component, then selected the first six components as the final input. In addition, Zhang, C. et al. [48] incorporated PCA into an AutoML framework for feature extraction, but did not demonstrate the optimal configuration.

An AutoEncoder consists of an encoder that maps the input x to a low-dimensional feature space z, and a decoder that reconstructs x from z. Khan, N. et al. [42] employed a convolutional AutoEncoder for extracting the features, improving the prediction accuracy for both residential and commercial energy consumption.

3.3.2. Feature Selection

Feature selection focuses on identifying the most relevant subset of features from the original dataset without generating new features. Feature selection has two main purposes: the first is to eliminate redundant features, and the second is to select important features.

Filter Selection

Filter selection is used to calculate some evaluation indicators that represent the correlation between the features or the importance of the features.

Olu-Ajayi, et al. [40] made a comparison among multiple feature selection techniques and pointed out that filter selection performs best overall in fault diagnosis. The inappropriate feature selection may have a negative effect on the ML model. ML algorithms such as XGBoost that do not rely on the number of features perform best without feature selection.

The Pearson coefficient measures linear correlation between two variables, whereas the Spearman coefficient measures monotonic correlation. Assadian, C.F. and Assadian, F. [41] calculated the Spearman coefficient between features and targets, and directly removed the feature of the non-monotonicity in building energy consumption. Zini and Carcasci [59] used both Pearson and Spearman coefficients across all features and the target variable (building electrical power).

Mutual Information Maximisation (MIM) is a filter technique that selects features that have the highest Mutual Information with the target. Ruan, Y. et al. [23] introduced the minimum Redundancy Maximum Relevance (mRMR) technique, which can consider the minimum redundancy on the basis of MIM. They compared both in terms of building electricity consumption and announced that mRMR is the better choice.

The symmetric uncertainty coefficient can also assess the correlation between variables. Mo and Zhao [47] chose a greedy strategy for searching the best feature combination and reduced the number of features a lot while keeping the performance of the model the same.

In addition, metrics such as Chi-square, Analysis of Variance, F-tests and Neighbourhood Component Analysis have also been employed [40,74]. Filter selections are always used to detect the redundant features by high correlation and identify the feature importance.

Embedded Selection

Embedded selection is to select the feature during the ML model training stage. The most used technique is the SelectKBest, which selects the top k most important features based on the evaluation metrics provided by the ML model. SelectKBest is concrete and effective, but it is limited to specific evaluation models and cannot consider the combined effects of features.

The ML algorithm Random Forest (RF) can be used with SelectKBest. Rahmanparast, A. et al. [22] analysed the feature importance indicator from the RF algorithm first and then validated and cross-checked the results using SelectKBest results. Assymkhan and Kartbayev [29] also used SelectKBest as a supplement to filter feature selection using RF. Different numbers of selected features were tested in their study. For the Support Vector Machine (SVM) and RF techniques, the suitable number of features is different.

Elastic Net, as an embedded selection technique, was used in the study of Saad, M.M. et al. [69]; it performed better than two other wrapper selection techniques on a multiple linear regression algorithm predicting the energy consumption, carbon emissions and cost. Similarly, Lian et al. [76] employed the embedded feature importance ranking mechanism of LightGBM to identify key variables influencing energy consumption prediction.

The SelectKBest is suitable for the ML model RF, and it is not always used independently but with filter selection. The application of embedded selection is limited; more techniques should be explored in the future.

Wrapper Selection

Wrapper selection is to select one or more ML algorithm models and use their performance on the test dataset as metrics. Wrapper selection considers the combined effects of features, but the disadvantage is the high training cost.

The technique recursive feature elimination (RFE) repeatedly builds models, removing the least important features each time until the number of features meets the requirements. Tian, J. et al. [33] used RFE with different ML algorithms for building electricity consumption prediction and reached a 10% improvement compared to the previous study, which was carried out on the same dataset. Zhang, C. et al. [48] used RFE with the XGBoost algorithm for building energy load forecasting, and the technique selected nearly all the features.

Qiao, Q. et al. [44] chose Boruta feature selection (BFS), whose core idea is to randomly copy the original feature set and then merge the copies with the chosen set to form an expanded set. They used BFS on teaching building and apartment hourly energy consumption prediction twice. The first BFS aims to exclude irrelevant features. The second BFS is implemented after the data decomposition process, focusing on generating the final feature set for energy consumption prediction.

In the study by Saad, M.M. et al. [69], both Backwards Stepwise Feature Selection (BSFS) and RFE are used to predict energy consumption, carbon emissions, and cost. BSFS took a longer time than RFE and had better performance on energy consumption and cost predictions.

According to the study of Olu-Ajayi et al. [40], wrapper selection is always the worst among different ML algorithms when used on fault detection and diagnosis. On the contrary, for the district heating system feature analysis, Weber, S.A. et al. [74] utilised wrapper selection based on the RNN ML algorithm. They found that the conclusion is similar to filter and embedded selections.

Knowledge-Based Selection

As was mentioned before, every feature in building operation data has a practical physical meaning. Therefore, researchers can use their expert knowledge to select features.

Researchers could first use filter selection, then they would take the coefficients as a reference and select appropriate features based on their expert knowledge. These applications have already been used for fault detection and diagnosis [54], heat flux prediction [60], and building energy consumption prediction [59].

Knowledge-based selection is always used in conjunction with filter selection to compensate for its poor interpretability. Some filter selection techniques are always based only on statistical characteristics of the dataset, so knowledge-based selection can also be a good way to involve expert knowledge in feature selection.

4. Results and Discussion

Table 4 compiles the data preprocessing techniques identified in the reviewed works, organised by steps, methods, and specific techniques.

Table 4. Summary of techniques.

Steps	Methods	Techniques	Application Scenery	References
Data Analysis	Exploratory Data Analysis	-	Widely recommended	[22–25,27,28,31–33,35,36,39,40,42,43,45,46,48,49,51–54,57–66,68,69,71–84]
	Clustering	Time-based Clustering	Optional (for categorising patterns)	[30,46,73,83]
		Algorithm Clustering	Optional (for categorising patterns)	[24,33,66,74,79,102]
	Temporal Statistical Test	-	Recommended for time-series regression	[75,94]
Data Preparation	Outlier Detection	-	Recommended for measuring data	[28,29,31,42,44,46,51,53,56,65,67,74,83]
	Missing Value Imputation	Statistical Imputation	Recommended for time-series data	[22,36,37,39,40,45,48,49,57,61,64,67,71,74,75,77,79,80]
		ML Model-based Imputation	Recommended for time-series data	[36,39,78,83]
	Data Transformation	Data Scaling	Recommended except the tree-based model	[24,26–28,30,35,37,42–46,50,52,60–62,65,66,69–71,73–75,77,79,80,84]
		Non-linear Transformations	Optional (for skewness correction)	[24,62]
		Data Smoothing	Optional (for removing noise)	[31,42,45,49,54,71,72]
		Data Decomposition	Optional (for splitting target value)	[23,43,44,48,61]
	Data Sampling	-	Recommended for imbalanced data	[53,55,56,69,76,83]
Feature Engineering	Feature Extraction	Encoding	Recommended	[29,34,36,39,41,42,44–46,50–52,57,59,63,66,67,72]
		Time Lag Features	Recommended for time-series regression	[32,46,58,63–65,67,68,75,80]
		Knowledge-based Features	Optional (expert knowledge)	[24,25,32,39,46,57,70,72,73,80]
		Dimension Reduction	Recommended for discrete datasets	[26,27,34,42,48,53]
	Feature Selection	Filter Selection	Optional (widely used)	[22,23,26,27,30,31,34,36,40,41,48,51–53,58–61,63,64,68,74,82]
		Embedded Selection	Optional (with specific ML models)	[22,29,40,58,74,76]
		Wrapper Selection	Optional	[33,40,44,48,58,63,69,74]
		Knowledge-based Selection	Optional (expert knowledge)	[54,60]

From the table, it can be inferred that the EDA of the data analysis step, data scaling for the data preparation step, encoding and filter selection for the feature engineering step have been widely used. Temporal statistical tests, non-linear transformation and knowledge-based feature selection have been little-explored.

Specifically for data analysis, EDA has been widely used, yet its proper name was not introduced, nor was a scientific procedure followed. Both time-based clustering and algorithm-based clustering can be used for office and commercial buildings, and the algorithms DBSCAN and K-means are the better options. The temporal statistical test is theoretically recommended for time-series modelling before time-series regression, but in the collected studies, it was reported in only a limited number of studies.

For the data preparation step, outlier detection is always achieved using IQR, while other algorithms might be useful. More comparative studies should be carried out. Statistical imputation is better for low-missing-rate data, while ML model-based imputation is better for high-missing-rate data. As for the data transformation method, data scaling is widely used by min-max scaling, and it performs well. Non-linear transformation is used when skewness is detected in the dataset. Among smoothing techniques, the Kalman filter was reported to achieve relatively stable performance in several studies. And STL and SSA serve as better data decomposition algorithms. Data sampling was mainly applied to address class imbalance in the FDD problem.

Feature extraction and feature selection form the step of feature engineering. In feature extraction, encoding, time-lag construction, physics-based feature creation, and dimensional reduction were discussed as techniques. Sine–cosine encoding effectively preserved the periodicity of time variables, while label encoding was suitable for ordinal levels such as fault severity. Time-lag features were sometimes chosen empirically. Physics-based features improved interpretability by integrating expert knowledge. PCA was widely applied for dimensionality reduction, while AutoEncoder showed growing potential. Feature selection methods included filter, embedded, wrapper, and knowledge-based selection techniques. Filter selection, especially correlation analysis, was the most common technique. Embedded selection, such as SelectKBest, is used as a supplement to filter selection, with the limitation of its reliance on the ML model. Wrapper selection could capture feature interactions but was associated with higher computational cost, and its performance varied across studies. Knowledge-based selection enhanced interpretability when combined with data-driven criteria, which is usually combined with filter selection.

In addition, a cross-dimensional examination of the selected studies reveals that preprocessing strategies are strongly influenced by data source, ML task type, and timestep.

First, the data source significantly affects the data preparation step. Measured datasets account for 69.4% of the reviewed studies, and missing value imputation appears in 19.4% of all studies, predominantly within measured data contexts. In contrast, simulation studies (22.6%) rarely report imputation or outlier detection. This indicates that data preparation is conditionally triggered by data quality rather than building type.

Regarding ML task type, regression and time-series regression together constitute approximately 82% of the studies. Data scaling with min–max normalisation appears in 37.1% of the studies and is mainly concentrated in regression contexts, confirming its role in stabilising continuous prediction models. Feature selection techniques are reported in 24.2% of the studies overall, and 30% of classification studies explicitly employ selection methods, suggesting that dimensional control is comparatively more visible in classification workflows.

In terms of timestep resolution, temporal feature engineering shows clear concentration in high-resolution settings. Time-lag features are used in 14.5% of the studies, and all such cases correspond to hourly prediction scenarios. Sine/cosine encoding appears in 9.7%

of the studies and is also restricted to hourly or sub-hourly time-series regression. These findings indicate that temporal dependency is primarily addressed through lag-based augmentation, while sine/cosine encoding remains selectively applied.

Overall, the cross-dimensional evidence demonstrates that the data source primarily drives data preparation, data transformation is closely linked to the regression task, and temporal feature engineering is activated almost exclusively under hourly or sub-hourly timesteps.

5. Conclusions

Based on the systematic review and analysis, the following conclusions address the proposed research questions.

- Applied techniques and workflows (RQ1):

This review proposed a three-step data preprocessing for ML-based building energy studies: data analysis, data preparation, and feature engineering. Each stage contains specific techniques that support effective model development. EDA, data scaling, outlier detection, feature encoding and filter feature selection were the most commonly used. However, EDA was frequently applied without a clearly structured methodological framework. Clustering, temporal stationarity tests, ML model-based imputation, non-linear transformation, data sampling and knowledge-based feature selection were less common. Techniques such as time-lag feature extraction and data decomposition showed inconsistent results across studies, indicating the need for further comparative research. Knowledge-based feature extraction and selection represent promising directions for integrating domain expertise.

Practically, preprocessing techniques should be selected according to data source, task type, data characteristics and ML models. The proposed workflow provides a structured guideline for preprocessing building energy data before ML model training.

- Strengths and weaknesses (RQ2):
  - (a) Data Analysis: The level of EDA varied considerably across studies. Public datasets were often described briefly, while private datasets showed more detail. Model-based clustering detects patterns but is difficult to interpret. In contrast, time-based clustering is easier to explain but difficult for detailed classification. Temporal stationarity tests, such as ADF and KPSS, were reported in a limited number of studies.
  - (b) Data Preparation: IQR was the most adopted outlier detection technique. Imputation techniques depend on missing rates: statistical techniques were used for missing rates below 30%, while model-based techniques were applied for higher missing rates. Scaling was mainly min–max normalisation, which is sensitive to extreme values. Kalman filtering shows stable performance but requires higher computational complexity. Decomposition studies show mixed performance: STL was stable, and VMD was powerful but demanding. Data sampling was applied primarily in two contexts: LHS expanded simulation data, while SMOTE helped with existing data imbalance.
  - (c) Feature Engineering: Sine/cosine encoding preserves the periodicity of time variables, and label encoding is good for the ordinal features. The time-lag feature can be extracted from analysis or experience. Physics-based features improved performance through expert knowledge. PCA was commonly applied for dimension reduction but was hard to interpret, while AutoEncoder has now gained wider use. Feature selection varied: statistical and embedded methods were reliable, wrapper methods costly, and expert-based selection risky.
- Gaps and misconceptions (RQ3):

Less than 20% of studies reported temporal stationarity tests. Most studies did not explicitly report data scaling or outlier detection procedures. Min–max scaling was sometimes wrongly assumed to handle outliers. Time features were sometimes encoded as ordinal integers, introducing artificial numerical relationships. Comparative studies of decomposition, imputation, and encoding methods are scarce.

In response to the above gaps, future research should focus on the following aspects: establishing a standardised preprocessing pipeline for ML applications in the building energy field. Preprocessing decisions, particularly scaling and encoding techniques, should be explicitly justified according to dataset characteristics and ML task type. More comparative studies of techniques on shared datasets should be proposed, using representative datasets to clearly indicate which preprocessing techniques are more suitable for datasets or tasks with specific characteristics.

This review has two main limitations. First, the scope of this review focuses only on the demand side, while the supply side could have different techniques to investigate. Second, the reviewed studies involve diverse datasets, ML tasks, and building types, so this review proposes a general methodological guidance rather than the optimal solution based on quantitative evidence.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/en19061561/s1, Table S1: PRISMA checklist.

Author Contributions: W.M.: Conceptualisation; Methodology; Data curation; Formal analysis; Investigation; Writing—original draft; Visualisation, Writing—review and editing. R.C.: Conceptualisation; Methodology; Supervision; Validation; Writing—review and editing. S.F.: Conceptualisation; Methodology; Supervision; Project administration; Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analysed in this study.

Conflicts of 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.

References

1. International Energy Agency. *Energy Efficiency 2024*; International Energy Agency: Paris, France, 2024.
2. Schito, E.; Conti, P.; Testi, D. Substitution of Heating Systems in the Italian Buildings Panorama and Potential for Energy, Environmental and Economic Efficiency Improvement. *Energy Build.* **2023**, *295*, 113273. [CrossRef]
3. Van Thillo, L.; Verbeke, S.; Audenaert, A. The Potential of Building Automation and Control Systems to Lower the Energy Demand in Residential Buildings: A Review of Their Performance and Influencing Parameters. *Renew. Sustain. Energy Rev.* **2022**, *158*, 112099. [CrossRef]
4. Li, D.; Qi, Z.; Zhou, Y.; Elchalakani, M. Machine Learning Applications in Building Energy Systems: Review and Prospects. *Buildings* **2025**, *15*, 648. [CrossRef]
5. Zhang, L.; Wen, J.; Li, Y.; Chen, J.; Ye, Y.; Fu, Y.; Livingood, W. A Review of Machine Learning in Building Load Prediction. *Appl. Energy* **2021**, *285*, 116452. [CrossRef]
6. Chen, Z.; O'Neill, Z.; Wen, J.; Pradhan, O.; Yang, T.; Lu, X.; Lin, G.; Miyata, S.; Lee, S.; Shen, C.; et al. A Review of Data-Driven Fault Detection and Diagnostics for Building HVAC Systems. *Appl. Energy* **2023**, *339*, 121030. [CrossRef]
7. Aguilar, J.; Garces-Jimenez, A.; R-Moreno, M.D.; García, R. A Systematic Literature Review on the Use of Artificial Intelligence in Energy Self-Management in Smart Buildings. *Renew. Sustain. Energy Rev.* **2021**, *151*, 111530. [CrossRef]
8. Jing, Q.; Guo, Y.; Liu, Y.; Wang, Y.; Du, C.; Liu, X. Optimization Study of Energy Saving Control Strategy of Carbon Dioxide Heat Pump Water Heater System under the Perspective of Energy Storage. *Appl. Therm. Eng.* **2026**, *283*, 129030. [CrossRef]
9. Das, H.P.; Lin, Y.-W.; Agwan, U.; Spangher, L.; Devonport, A.; Yang, Y.; Drgoňa, J.; Chong, A.; Schiavon, S.; Spanos, C.J. Machine Learning for Smart and Energy-Efficient Buildings. *Environ. Data Sci.* **2024**, *3*, e1. [CrossRef]
10. Getting Started with Data-Centric AI Development: Tips from Andrew Ng. Available online: https://www.elucidata.io/blog/getting-started-with-data-centric-ai-development (accessed on 3 June 2025).

11. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
12. Fan, C.; Chen, M.; Wang, X.; Wang, J.; Huang, B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data. *Front. Energy Res.* **2021**, *9*, 652801. [CrossRef]
13. Wang, Z.; Liu, J.; Zhang, Y.; Yuan, H.; Zhang, R.; Srinivasan, R.S. Practical Issues in Implementing Machine-Learning Models for Building Energy Efficiency: Moving beyond Obstacles. *Renew. Sustain. Energy Rev.* **2021**, *143*, 110929. [CrossRef]
14. Sun, Y.; Haghighat, F.; Fung, B.C.M. A Review of The-State-of-the-Art in Data-Driven Approaches for Building Energy Prediction. *Energy Build.* **2020**, *221*, 110022. [CrossRef]
15. Hou, D.; Evins, R. A Protocol for Developing and Evaluating Neural Network-Based Surrogate Models and Its Application to Building Energy Prediction. *Renew. Sustain. Energy Rev.* **2024**, *193*, 114283. [CrossRef]
16. Zhang, Y.; Wang, D.; Wang, G.; Xu, P.; Zhu, Y. Data-Driven Building Load Prediction and Large Language Models: Comprehensive Overview. *Energy Build.* **2025**, *326*, 115001. [CrossRef]
17. Liu, H.; Liang, J.; Liu, Y.; Wu, H. A Review of Data-Driven Building Energy Prediction. *Buildings* **2023**, *13*, 532. [CrossRef]
18. Chen, Y.; Gong, W.; Obrecht, C.; Kuznik, F. A Review of Machine Learning Techniques for Building Electrical Energy Consumption Prediction. *Energy AI* **2025**, *21*, 100518. [CrossRef]
19. Chen, G.; Lu, S.; Zhou, S.; Tian, Z.; Kim, M.K.; Liu, J.; Liu, X. A Systematic Review of Building Energy Consumption Prediction: From Perspectives of Load Classification, Data-Driven Frameworks, and Future Directions. *Appl. Sci.* **2025**, *15*, 3086. [CrossRef]
20. Dai, X.; Liu, J.; Zhang, X. A Review of Studies Applying Machine Learning Models to Predict Occupancy and Window-Opening Behaviours in Smart Buildings. *Energy Build.* **2020**, *223*, 110159. [CrossRef]
21. Wang, Z.; Xia, L.; Yuan, H.; Srinivasan, R.S.; Song, X. Principles, Research Status, and Prospects of Feature Engineering for Data-Driven Building Energy Prediction: A Comprehensive Review. *J. Build. Eng.* **2022**, *58*, 105028. [CrossRef]
22. Rahmanparast, A.; Milani, M.; Camci, M.; Karakoyun, Y.; Acikgoz, O.; Dalkilic, A.S. A Comprehensive Method for Exploratory Data Analysis and Preprocessing the ASHRAE Database for Machine Learning. *Appl. Therm. Eng.* **2025**, *273*, 126556. [CrossRef]
23. Ruan, Y.; Wang, G.; Meng, H.; Qian, F. A Hybrid Model for Power Consumption Forecasting Using VMD-Based the Long Short-Term Memory Neural Network. *Front. Energy Res.* **2022**, *9*, 772508. [CrossRef]
24. Klemp, S.; Abida, A.; Richter, P. A Method and Analysis of Predicting Building Material U-Value Ranges through Geometrical Pattern Clustering. *J. Build. Eng.* **2021**, *44*, 103243. [CrossRef]
25. Jeong, J.; Lee, D.; Chae, Y.T. A Novel Approach for Day-Ahead Hourly Building-Integrated Photovoltaic Power Prediction by Using Feature Engineering and Simple Weather Forecasting Service. *Energies* **2023**, *16*, 7477. [CrossRef]
26. Kumar Mohapatra, S.; Mishra, S.; Tripathy, H.K.; Alkhayyat, A. A Sustainable Data-Driven Energy Consumption Assessment Model for Building Infrastructures in Resource Constraint Environment. *Sustain. Energy Technol. Assess.* **2022**, *53*, 102697. [CrossRef]
27. Ramana, B.S.S.V.; Kumar, S.C.; Bharath Kumar, N.; Babu, A.R.V. A Web-Based Machine Learning Framework for Building Energy Efficiency Prediction. *Frankl. Open* **2025**, *11*, 100284. [CrossRef]
28. Mahmood, S.; Sun, H.; Ali Alhussan, A.; Iqbal, A.; El-kenawy, E.-S.M. Active Learning-Based Machine Learning Approach for Enhancing Environmental Sustainability in Green Building Energy Consumption. *Sci. Rep.* **2024**, *14*, 19894. [CrossRef]
29. Assymkhan, N.; Kartbayev, A. Advanced IoT-Enabled Indoor Thermal Comfort Prediction Using SVM and Random Forest Models. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 1040–1050. [CrossRef]
30. Jeevakarunya, C.; Manikandan, V. Advanced Machine Learning Approach with Dynamic Kernel Weighting for Accurate Electrical Load Forecasting. *AIP Adv.* **2025**, *15*, 015011. [CrossRef]
31. Es-sakali, N.; Zoubir, Z.; Idrissi Kaitouni, S.; Mghazli, M.O.; Cherkaoui, M.; Pfafferott, J. Advanced Predictive Maintenance and Fault Diagnosis Strategy for Enhanced HVAC Efficiency in Buildings. *Appl. Therm. Eng.* **2024**, *254*, 123910. [CrossRef]
32. Moon, J.; Maqsood, M.; So, D.; Baik, S.W.; Rho, S.; Nam, Y. Advancing Ensemble Learning Techniques for Residential Building Electricity Consumption Forecasting: Insight from Explainable Artificial Intelligence. *PLoS ONE* **2024**, *19*, e0307654. [CrossRef]
33. Tian, J.; Li, K.; Xue, W. An Adaptive Ensemble Predictive Strategy for Multiple Scale Electrical Energy Usages Forecasting. *Sustain. Cities Soc.* **2021**, *66*, 102654. [CrossRef]
34. Durand, D.; Aguilar, J.; R-Moreno, M.D. An Analysis of the Energy Consumption Forecasting Problem in Smart Buildings Using LSTM. *Sustainability* **2022**, *14*, 13358. [CrossRef]
35. Sadeghi, A.; Sinaki, R.Y.; Young, W.A.; Weckman, G.R. An Intelligent Model to Predict Energy Performances of Residential Buildings Based on Deep Neural Networks. *Energies* **2020**, *13*, 571. [CrossRef]
36. Hussain, A.; Giangrande, P.; Franchini, G.; Fenili, L.; Messi, S. Analyzing the Effect of Error Estimation on Random Missing Data Patterns in Mid-Term Electrical Forecasting. *Electronics* **2025**, *14*, 1383. [CrossRef]
37. Pydi, D.P.; Advaith, S. Attention Boosted Autoencoder for Building Energy Anomaly Detection. *Energy AI* **2023**, *14*, 100292. [CrossRef]

38. Hosseini Gourabpasi, A.; Nik-Bakht, M. BIM-Based Automated Fault Detection and Diagnostics of HVAC Systems in Commercial Buildings. *J. Build. Eng.* **2024**, *87*, 109022. [CrossRef]
39. Buțurache, A.-N.; Stancu, S. Building Energy Consumption Prediction Using Neural-Based Models. *Int. J. Energy Econ. Policy* **2022**, *12*, 30–38. [CrossRef]
40. Olu-Ajayi, R.; Alaka, H.; Sulaimon, I.; Balogun, H.; Wusu, G.; Yusuf, W.; Adegoke, M. Building Energy Performance Prediction: A Reliability Analysis and Evaluation of Feature Selection Methods. *Expert Syst. Appl.* **2023**, *225*, 120109. [CrossRef]
41. Assadian, C.F.; Assadian, F. Data-Driven Modeling of Appliance Energy Usage. *Energies* **2023**, *16*, 7536. [CrossRef]
42. Khan, N.; Khan, S.U.; Baik, S.W. Deep Autoencoder-Based Hybrid Network for Building Energy Consumption Forecasting. *Comput. Syst. Sci. Eng.* **2024**, *48*, 154–173. [CrossRef]
43. de Rautlin de la Roy, E.; Recht, T.; Zemmari, A.; Bourreau, P.; Mora, L. Deep Learning Models for Building Window-Openings Detection in Heating Season. *Build. Environ.* **2023**, *231*, 110019. [CrossRef]
44. Qiao, Q.; Yunusa-Kaltungo, A.; Edwards, R.E. Developing a Machine Learning Based Building Energy Consumption Prediction Approach Using Limited Data: Boruta Feature Selection and Empirical Mode Decomposition. *Energy Rep.* **2023**, *9*, 3643–3660. [CrossRef]
45. Ullah, F.U.M.; Khan, N.; Hussain, T.; Lee, M.Y.; Baik, S.W. Diving Deep into Short-term Electricity Load Forecasting: Comparative Analysis and a Novel Framework. *Mathematics* **2021**, *9*, 611. [CrossRef]
46. Liu, Y.; Zhao, X.; Qin, S.J. Dynamically Engineered Multi-Modal Feature Learning for Predictions of Office Building Cooling Loads. *Appl. Energy* **2024**, *355*, 122183. [CrossRef]
47. Mo, Y.; Zhao, D. Effective Factors for Residential Building Energy Modeling Using Feature Engineering. *J. Build. Eng.* **2021**, *44*, 102891. [CrossRef]
48. Zhang, C.; Lu, J.; Huang, J.; Zhao, Y. End-to-End Data-Driven Modeling Framework for Automated and Trustworthy Short-Term Building Energy Load Forecasting. *Build. Simul.* **2024**, *17*, 1419–1437. [CrossRef]
49. Nazir, K.; Memon, S.A. Evaluating the Impact of Data Preprocessing to Develop a Robust MEP-Based Forecasting Model for Building Integrated with PCM. *Energy* **2025**, *324*, 135763. [CrossRef]
50. Sayed, A.N.; Himeur, Y.; Bensaali, F. From Time-Series to 2D Images for Building Occupancy Prediction Using Deep Transfer Learning. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105786. [CrossRef]
51. Perez Garcia, C.A.; Bovo, M.; Torreggiani, D.; Tassinari, P.; Benni, S. Indoor Temperature Forecasting in Livestock Buildings: A Data-Driven Approach. *Agriculture* **2024**, *14*, 316. [CrossRef]
52. Mahmood, S.; Sun, H.; El-kenawy, E.-S.M.; Iqbal, A.; Alharbi, A.H.; Khafaga, D.S. Integrating Machine and Deep Learning Technologies in Green Buildings for Enhanced Energy Efficiency and Environmental Sustainability. *Sci. Rep.* **2024**, *14*, 20331. [CrossRef]
53. Asadi, N.; Moosavi, L. Investigation of Window Opening Behavior during Cold Seasons through a Non-Intrusive Sensor-Based Data-Driven Approach. *Energy Build.* **2024**, *317*, 114386. [CrossRef]
54. Du, Z.; Chen, S.; Anduv, B.; Zhu, X.; Jin, X. IoT Intelligent Agent Based Cloud Management System by Integrating Machine Learning Algorithm for HVAC Systems. *Int. J. Refrig.* **2023**, *146*, 158–173. [CrossRef]
55. Mounter, W.; Ogwumike, C.; Dawood, H.; Dawood, N. Machine Learning and Data Segmentation for Building Energy Use Prediction—A Comparative Study. *Energies* **2021**, *14*, 5947. [CrossRef]
56. Chen, Y.; Ye, Y.; Liu, J.; Zhang, L.; Li, W.; Mohtaram, S. Machine Learning Approach to Predict Building Thermal Load Considering Feature Variable Dimensions: An Office Building Case Study. *Buildings* **2023**, *13*, 312. [CrossRef]
57. Mansouri, A.; Naghdi, M.; Erfani, A. Machine Learning for Leadership in Energy and Environmental Design Credit Targeting: Project Attributes and Climate Analysis Toward Sustainability. *Sustainability* **2025**, *17*, 2521. [CrossRef]
58. Alkhulaifi, N.; Bowler, A.L.; Pekaslan, D.; Serdaroglu, G.; Closs, S.; Watson, N.J.; Triguero, I. Machine Learning Pipeline for Energy and Environmental Prediction in Cold Storage Facilities. *IEEE Access* **2024**, *12*, 153935–153951. [CrossRef]
59. Zini, M.; Carcasci, C. Machine Learning-Based Monitoring Method for the Electricity Consumption of a Healthcare Facility in Italy. *Energy* **2023**, *262*, 125576. [CrossRef]
60. Shen, Z.; Shrestha, S.; Howard, D.; Feng, T.; Hun, D.; She, B. Machine Learning–Assisted Prediction of Heat Fluxes through Thermally Anisotropic Building Envelopes. *Build. Environ.* **2023**, *234*, 110157. [CrossRef]
61. Fernández-Martínez, D.; Jaramillo-Morán, M.A. Multi-Step Hourly Power Consumption Forecasting in a Healthcare Building with Recurrent Neural Networks and Empirical Mode Decomposition. *Sensors* **2022**, *22*, 3664. [CrossRef]
62. Yang, N.-C.; Sung, K.-L. Non-Intrusive Load Classification and Recognition Using Soft-Voting Ensemble Learning Algorithm with Decision Tree, K-Nearest Neighbor Algorithm and Multilayer Perceptron. *IEEE Access* **2023**, *11*, 94506–94520. [CrossRef]
63. Kim, J.; Frank, S.; Buechler, R.; Mishra, S.; Petersen, A.; Zhang, L.; Eslinger, H. Performance Evaluation of Automated Data-Driven Feature Extraction and Selection Methods for Practical and Scalable Building Energy Consumption Prediction Models. *J. Build. Eng.* **2025**, *103*, 112045. [CrossRef]

64. Attipoe, D.; Moulla, D.K.; Mnkandla, E.; Abran, A. Predicting Residential Energy Consumption in South Africa Using Ensemble Models. *Appl. Comput. Intell. Soft Comput.* **2025**, *2025*, 5211419. [CrossRef]
65. Moulla, D.K.; Attipoe, D.; Mnkandla, E.; Abran, A. Predictive Model of Energy Consumption Using Machine Learning: A Case Study of Residential Buildings in South Africa. *Sustainability* **2024**, *16*, 4365. [CrossRef]
66. Tang, L.; Xie, H.; Wang, X.; Bie, Z. Privacy-Preserving Knowledge Sharing for Few-Shot Building Energy Prediction: A Federated Learning Approach. *Appl. Energy* **2023**, *337*, 120860. [CrossRef]
67. Moon, J.; Park, S.; Rho, S.; Hwang, E. Robust Building Energy Consumption Forecasting Using an Online Learning Approach with R Ranger. *J. Build. Eng.* **2022**, *47*, 103851. [CrossRef]
68. Alrashidi, M. Short-Term Mosques Load Forecast Using Machine Learning and Meteorological Data. *Comput. Syst. Sci. Eng.* **2023**, *46*, 371–387. [CrossRef]
69. Saad, M.M.; Menon, R.P.; Eicker, U. Supporting Decision Making for Building Decarbonization: Developing Surrogate Models for Multi-Criteria Building Retrofitting Analysis. *Energies* **2023**, *16*, 6030. [CrossRef]
70. Sajjad, M.; Khan, S.U.; Khan, N.; Haq, I.U.; Ullah, A.; Lee, M.Y.; Baik, S.W. Towards Efficient Building Designing: Heating and Cooling Load Prediction via Multi-Output Model. *Sensors* **2020**, *20*, 6419. [CrossRef]
71. Khan, S.U.; Iqbal, E.; Khan, N.; Zweiri, Y.; Abdulrahman, Y. Towards Net Zero Energy Building: AI-Based Framework for Power Consumption and Generation Prediction. *Energy Build.* **2025**, *331*, 115311. [CrossRef]
72. Wang, H.; Pawlak, J.; Faghih Imani, A.; Guo, F.; Sivakumar, A. When Does It Pay off to Use Electricity Demand Data with Rich Information about Households and Their Activities? A Comparative Machine Learning Approach to Demand Modelling. *Energy Build.* **2023**, *295*, 113292. [CrossRef]
73. Almadhor, A.; Alsubai, S.; Kryvinska, N.; Ghazouani, N.; Bouallegue, B.; Al Hejaili, A.; Sampedro, G.A. A Synergistic Approach Using Digital Twins and Statistical Machine Learning for Intelligent Residential Energy Modelling. *Sci. Rep.* **2025**, *15*, 26088. [CrossRef]
74. Weber, S.A.; Fischlschweiger, M.; Volta, D.; Geisler, J. Feature Selection for Specific Prediction Targets at the User Level in a District Heating Network. *Sci. Rep.* **2025**, *15*, 29789. [CrossRef] [PubMed]
75. Pai H, A.; Mishra, K.K.; Mahesh, T.R.; Jeyan, J.V.M.L.; Sayal, A. Enhanced Household Energy Consumption Forecasting Using Multivariate Long Short-Term Memory (LSTM) Networks with Weather Data Integration. *Results Eng.* **2025**, *27*, 106512. [CrossRef]
76. Lian, H.; Ji, Y.; Niu, M.; Gu, J.; Xie, J.; Liu, J. A Hybrid Load Prediction Method of Office Buildings Based on Physical Simulation Database and LightGBM Algorithm. *Appl. Energy* **2025**, *377*, 124620. [CrossRef]
77. Telicko, J.; Jakovics, A. Applying Dynamic U-Value Measurements for State Forecasting in Buildings. *Latv. J. Phys. Tech. Sci.* **2023**, *60*, 81–94. [CrossRef]
78. Muslimsyah, M.; Safwan, S.; Novandri, A. Comprehensive Assessment of Indoor Thermal in Vernacular Building Using Machine Learning Model with GAN-Based Data Imputation: A Case of Aceh Region, Indonesia. *Buildings* **2025**, *15*, 2448. [CrossRef]
79. Kim, D.; Seomun, G.; Lee, Y.; Cho, H.; Chin, K.; Kim, M.-H. Forecasting Building Energy Demand and On-Site Power Generation for Residential Buildings Using Long and Short-Term Memory Method with Transfer Learning. *Appl. Energy* **2024**, *368*, 123500. [CrossRef]
80. Papias, I.; Michalakopoulos, V.; Sarmas, E.; Marinakis, V.; Antonesi, G.; Cioara, T.; Anghel, I. A Data-Driven Framework for Estimating Residential Energy Flexibility for Aggregated Demand-Side Management. *Sustain. Energy Grids Netw.* **2025**, *43*, 101783. [CrossRef]
81. Alotaibi, B.S. Advancing Energy Performance Efficiency in Residential Buildings for Sustainable Design: Integrating Machine Learning and Optimized Explainable AI (AIX). *Int. J. Energy Res.* **2024**, *2024*, 6130634. [CrossRef]
82. Fellah, M.; Ouhaibi, S.; Belouaggadia, N.; Mansouri, K. Energy Consumption Forecasting and Thermal Insulator Selection with Random Forest Regression. *Sci. Afr.* **2025**, *29*, e02870. [CrossRef]
83. Bandória, L.H.T.; de Almeida, M.C. Modeling and Estimating Standard Deviation of Active Power Demand Using a Multi-Model Stacking Regressor. *Sustain. Energy Grids Netw.* **2025**, *43*, 101788. [CrossRef]
84. Pintea, A. Sensor-Based Room Inhabitance Monitoring Using Robust ML Models Compatible with Large Datasets/Real-Time Datastreams. *J. Univers. Comput. Sci.* **2025**, *31*, 1665–1689. [CrossRef]
85. Ucar, M.T.; Kaygusuz, A. Short-Term Energy Consumption Forecasting Analysis Using Different Optimization and Activation Functions with Deep Learning Models. *Appl. Sci.* **2025**, *15*, 6839. [CrossRef]
86. Javed, U.; Ijaz, K.; Jawad, M.; Ansari, E.A.; Shabbir, N.; Kütt, L.; Husev, O. Exploratory Data Analysis Based Short-Term Electrical Load Forecasting: A Comprehensive Analysis. *Energies* **2021**, *14*, 5510. [CrossRef]
87. Yang, Q.; Lu, W.; Xu, F.; Luo, X.; Wen, B. A Passive Strategy for Energy-Saving Retrofitting of Courtyard Dwellings and Its Climatic Adaptability. *Energy Build.* **2026**, *352*, 116811. [CrossRef]
88. Yao, G.; Guo, C.; Ge, Q.; Ait-Ahmed, M. A Practical Building Energy Consumption Anomaly Detection Method Based on Parameter Adaptive Setting DBSCAN. *Cogn. Comput. Syst.* **2021**, *3*, 154–168. [CrossRef]

89. Arias-Requejo, D.; Pulido, B.; Keane, M.M.; Alonso-González, C.J. Clustering and Deep-Learning for Energy Consumption Forecast in Smart Buildings. *IEEE Access* **2023**, *11*, 128061–128080. [CrossRef]
90. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]
91. Cui, Y.; Zhu, Z.; Zhao, X.; Li, Z. Energy Schedule Setting Based on Clustering Algorithm and Pattern Recognition for Non-Residential Buildings Electricity Energy Consumption. *Sustainability* **2023**, *15*, 8750. [CrossRef]
92. Liu, K.; Wang, X.; Xue, L. Circuit Categorization Approach of Office Building Energy Consumption Based on Data Features for Energy-Saving Diagnosis. *Energy Build.* **2024**, *323*, 114811. [CrossRef]
93. Benitez, I.B.; Ibañez, J.A.; Lumabad, C.I.D.; Cañete, J.M.; Principe, J.A. Day-Ahead Hourly Solar Photovoltaic Output Forecasting Using SARIMAX, Long Short-Term Memory, and Extreme Gradient Boosting: Case of the Philippines. *Energies* **2023**, *16*, 7823. [CrossRef]
94. Chou, S.-Y.; Dewabharata, A.; Zulvia, F.E.; Fadil, M. Forecasting Building Energy Consumption Using Ensemble Empirical Mode Decomposition, Wavelet Transformation, and Long Short-Term Memory Algorithms. *Energies* **2022**, *15*, 1035. [CrossRef]
95. Ma, J.; Cheng, J.C.P.; Jiang, F.; Chen, W.; Wang, M.; Zhai, C. A Bi-Directional Missing Data Imputation Scheme Based on LSTM and Transfer Learning for Building Energy Data. *Energy Build.* **2020**, *216*, 109941. [CrossRef]
96. Zhang, L. A Pattern-Recognition-Based Ensemble Data Imputation Framework for Sensors from Building Energy Systems. *Sensors* **2020**, *20*, 5947. [CrossRef] [PubMed]
97. Cho, B.; Dayrit, T.; Gao, Y.; Wang, Z.; Hong, T.; Sim, A.; Wu, K. Effective Missing Value Imputation Methods for Building Monitoring Data. In *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), December 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 2866–2875.
98. de Amorim, L.B.V.; Cavalcanti, G.D.C.; Cruz, R.M.O. The Choice of Scaling Technique Matters for Classification Performance. *Appl. Soft Comput.* **2023**, *133*, 109924. [CrossRef]
99. Gao, Y.; Ruan, Y.; Fang, C.; Yin, S. Deep Learning and Transfer Learning Models of Energy Consumption Forecasting for a Building with Poor Information Data. *Energy Build.* **2020**, *223*, 110156. [CrossRef]
100. Wen, S.; Zhang, W.; Sun, Y.; Li, Z.; Huang, B.; Bian, S.; Zhao, L.; Wang, Y. An Enhanced Principal Component Analysis Method with Savitzky–Golay Filter and Clustering Algorithm for Sensor Fault Detection and Diagnosis. *Appl. Energy* **2023**, *337*, 120862. [CrossRef]
101. Kim, D.-W.; Park, C.-S. Application of Kalman Filter for Estimating a Process Disturbance in a Building Space. *Sustainability* **2017**, *9*, 1868. [CrossRef]
102. El Assri, N.; Jallal, M.A.; Chabaa, S.; Zeroual, A. Enhancing Building Energy Consumption Prediction Using LSTM, Kalman Filter, and Continuous Wavelet Transform. *Sci. Afr.* **2025**, *27*, e02560. [CrossRef]
103. Zhang, X.; Li, R. A Novel Decomposition and Combination Technique for Forecasting Monthly Electricity Consumption. *Front. Energy Res.* **2021**, *9*, 792358. [CrossRef]
104. Wei, S.; Bai, X. Multi-Step Short-Term Building Energy Consumption Forecasting Based on Singular Spectrum Analysis and Hybrid Neural Network. *Energies* **2022**, *15*, 1743. [CrossRef]
105. Liu, J.; Lv, Z.; Zhao, L. A Dual-Optimization Building Energy Prediction Framework Based on Improved Dung Beetle Algorithm, Variational Mode Decomposition and Deep Learning. *Energy Build.* **2025**, *328*, 115143. [CrossRef]
106. Jin, N.; Yang, F.; Mo, Y.; Zeng, Y.; Zhou, X.; Yan, K.; Ma, X. Highly Accurate Energy Consumption Forecasting Model Based on Parallel LSTM Neural Networks. *Adv. Eng. Inform.* **2022**, *51*, 101442. [CrossRef]
107. Ali, U.; Bano, S.; Shamsi, M.H.; Sood, D.; Hoare, C.; Zuo, W.; Hewitt, N.; O'Donnell, J. Urban Building Energy Performance Prediction and Retrofit Analysis Using Data-Driven Machine Learning Approach. *Energy Build.* **2023**, *303*, 113768. [CrossRef]
108. Xu, W.; Wu, X.; Xiong, S.; Li, T.; Liu, Y. Optimizing the Sustainable Performance of Public Buildings: A Hybrid Machine Learning Algorithm. *Energy* **2025**, *320*, 135283. [CrossRef]
109. Yang, Q.; Xu, F.; Lu, W.; Yang, Z.; Bai, Y.; Wen, B. Green Renovation and Multi-Objective Optimization of Tibetan Courtyard Dwellings. *Build. Environ.* **2025**, *279*, 113071. [CrossRef]
110. Wang, Y.; Li, Z.; Chen, H.; Zhang, J.; Liu, Q.; Wu, J.; Shen, L. Research on Diagnostic Strategy for Faults in VRF Air Conditioning System Using Hybrid Data Mining Methods. *Energy Build.* **2021**, *247*, 111144. [CrossRef]
111. Ci, T.; Liu, Z.; Wang, Y. Assessment of the Degree of Building Damage Caused by Disaster Using Convolutional Neural Networks in Combination with Ordinal Regression. *Remote Sens.* **2019**, *11*, 2858. [CrossRef]
112. Guo, Y.; Du, C.; Liu, X.; Zhang, X.; Jin, Z. Research on Attention-Based Fault Diagnosis and Multi-Parameter Joint Optimization of CO2 Heat Pump System. *Appl. Therm. Eng.* **2026**, *289*, 129942. [CrossRef]
113. Zhang, L.; Plathottam, S.; Reyna, J.; Merket, N.; Sayers, K.; Yang, X.; Reynolds, M.; Parker, A.; Wilson, E.; Fontanini, A.; et al. High-Resolution Hourly Surrogate Modeling Framework for Physics-Based Large-Scale Building Stock Modeling. *Sustain. Cities Soc.* **2021**, *75*, 103292. [CrossRef]

114. Xie, J.; Chen, Y.; Hong, T.; Laing, T.D. Relative Humidity for Load Forecasting Models. *IEEE Trans. Smart Grid* **2018**, *9*, 191–198. [CrossRef]
115. Xie, J.; Hong, T. Wind Speed for Load Forecasting Models. *Sustainability* **2017**, *9*, 795. [CrossRef]