

Article

Driver–Pathway Analysis of EUI in Historic Buildings: Rank Fusion and Rolling Validation

Chen Liu ¹, Fuying Liu ^{2,*} and Qi Zhao ³

¹ School of Resources and Civil Engineering, Northeastern University, Shenyang 110819, China; liuc53@mails.neu.edu.cn

² JangHo Architecture College, Northeastern University, Shenyang 110819, China

³ School of Management, Shenyang Jianzhu University, Shenyang 110168, China; dbdxzq@163.com

* Correspondence: lfycl@sina.com

Abstract

Historic buildings often exhibit high energy use intensity (EUI), while conservation constraints limit envelope retrofits, making it difficult to identify robust and actionable operational predictors. Using four in-use historic buildings in Shenyang, China, this study presents a pilot methodological demonstration with a controlled-comparability workflow consisting of two linked layers: (i) a Driver layer of intervenable operational variables and (ii) a Pathway layer of calibrated EnergyPlus heat-balance terms for physics-informed interpretation. Three importance approaches (Spearman, wrapper RFE with XGBoost, and Random Forest) are compared; rankings are fused via reciprocal rank fusion, and stability is tested using cross-period rolling validation across Top-K feature sets. After similarity screening, EUI variation is better explained by operational predictors and the corresponding simulated loss channels than by macro-scale structural heterogeneity. Infiltration-related indicators and envelope/infiltration loss components remain consistently prominent, while Spearman importance is less stable in the Pathway layer under seasonal switching and nonlinear coupling. A Top-10 subset provides a favorable accuracy–stability trade-off. The proposed Driver–Pathway mapping supports conservation-compatible prioritization hypotheses within a simulation-consistent interpretive framework; findings are associational and context dependent and should be validated through field measurements and experimental or quasi-experimental studies before prescriptive claims are made.

Keywords: building energy consumption; historic buildings; feature importance; building energy efficiency; energy use intensity (EUI); SHAP



Academic Editor: Moncef Krarti

Received: 19 February 2026

Revised: 26 March 2026

Accepted: 29 March 2026

Published: 7 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Improving energy efficiency in the building sector is a priority for the construction industry. In urban agglomerations, total energy demand can reach up to 80% of the global level, and buildings alone contribute roughly 40% of that consumption [1]. Across different countries and regions, buildings represent a large share of overall energy use, although the proportion varies substantially—for instance, 23% in Spain, 25% in Japan, 28% in China, 39% in the United Kingdom, about 40% in Europe, 42% in Brazil, 47% in Switzerland, and 50% in Botswana [2–4]. This prominence is partly explained by occupancy patterns: people spend more than 65% of their time indoors [5]. As a result, building energy use has attracted sustained attention worldwide [6]. In China, emissions from building operations reached 2.16 billion metric tons in 2020, accounting for approximately 22% of national

emissions, and this fraction may increase in the coming years [7,8]. In the same year, the Chinese government announced the “30–60” targets for carbon peaking and carbon neutrality; corresponding actions to peak building-related emissions are expected to be advanced progressively through city-level implementation [9]. As can be seen, reducing building energy consumption or improving the energy efficiency of buildings is a problem that is highly valued by China and urgently needs to be addressed.

Historic buildings form a distinct segment within the existing building stock, and their scale is substantial in many regions. In the European Union, about 35% of buildings are older than 50 years, and nearly three quarters are considered energy-inefficient [10]. For example, the 2019 census indicates that roughly two million buildings in Italy are more than 100 years old [11]. In the United Kingdom, it is estimated that around 20% of buildings may possess heritage value [12]. Buildings constructed before 1920 contribute approximately 22% of the final energy consumption of residences [13]. China also has a large inventory of historic buildings [14]. Official statistics report 67,200 designated historic buildings nationwide [15]. There are a vast number of historic buildings worldwide, and many of these historic buildings are still in use and their energy efficiency performance is relatively poor. Enhancing the energy performance of this building category can therefore support reductions in overall energy demand and greenhouse-gas emissions. Furthermore, historic buildings are symbols of a country’s history, culture and national spirit, and they are important signs of a nation’s identity. Maintaining historic buildings in optimal condition constitutes a significant undertaking and holds substantial value for a nation. A practical way to improve outcomes while avoiding neglect or abandonment is to keep these buildings in active use through rehabilitation and adaptive reuse [16]. By updating functions to meet contemporary needs, such projects can preserve architectural character and historical significance, while contributing to more sustainable and resilient cities [17,18].

Adaptive reuse of historic buildings has been widely endorsed from multiple perspectives. From a policy and heritage-governance standpoint, several international documents stress that conservation should integrate cultural significance with continued use, because reuse helps sustain the vitality of heritage assets. This principle is reflected in the Nara Document on Authenticity (1994) [19] and the Burra Charter (2013) [16]. The Faro Convention (2005) further highlights the importance of enabling historic buildings to serve contemporary societal needs through comprehensive reuse [20]. From a sustainability perspective, adaptive reuse also aligns with circular-economy thinking, which emphasizes closed-loop resource use [21]. Retaining and upgrading existing historic buildings can reduce demand for virgin materials and avoid the resource intensity associated with demolition and new construction [22]. By extending service life, reuse can generate broader co-benefits beyond the building itself, including positive impacts on social development, environmental protection, and economic growth [23]. In addition, reuse can contribute to lowering operational energy demand, which is essential in responding to energy challenges [24]. However, maintaining functionality after reuse requires ongoing monitoring and maintenance. Prior studies have therefore explored sensor-based approaches to capture real-time information on indoor environmental conditions, surface states [25], and structural health [26]. With advances in cloud computing, general-purpose frameworks have been developed to support indoor environmental quality monitoring and management at scale [27,28]. In particular, public cloud services can improve system reliability and scalability while helping reduce deployment costs.

Recent advances in data-centric methods—particularly machine learning and artificial intelligence—have reshaped research on historic-building energy performance. This research has evolved from initially relying on sensors for indoor environment and building monitoring to incorporating cloud computing and other platform-based frameworks to

support data collection, storage, and management. The focus has expanded from the primary goal of protecting historic buildings to providing decision support for energy-saving renovations without compromising their historic value and structural safety, aiming to achieve the coordinated goal of protection and energy conservation. In recent years, researchers have increasingly adopted methods such as sensitivity analysis, machine learning, and deep learning to model energy use in historic buildings and analyze influencing factors. Rather than relying solely on sensor-based monitoring, current studies integrate data-driven workflows to identify key predictors and potential weak links, thereby informing more targeted, conservation-aware retrofit hypotheses and evaluation priorities. These methods have become one of the important research paths in this field. Historic buildings are different from other buildings. They also need to take other aspects into consideration, including building age, the thermal and physical characteristics of traditional materials [29], and building function or usage patterns. Meanwhile, building energy simulation has long been used to pre-evaluate the feasibility and expected effectiveness of retrofit measures [30]. More recently, simulation has also been applied to examine intervention impacts from a conservation-oriented perspective [31,32]. For historic buildings, given their significant historic value, scientific assessment is necessary before any renovation to ensure that such intervention measures do not damage the integrity of the buildings [15]. Accurately assessing building energy performance is challenging because building energy-use behavior is heterogeneous and influenced by many interacting variables [33]. Even when the same set of attributes is considered, their effects on energy use can vary substantially across different buildings. Therefore, selecting appropriate features typically requires a solid knowledge of building physics as well as a detailed understanding of the specific building under study [34]. Therefore, given that building energy consumption simulation is influenced by various types and numerous factors, it is crucial to select the influencing factors and take them into account, especially for historic buildings.

1.1. Previous Studies

Researchers use building energy consumption simulation method to study historic energy consumption problems. As an example, Fabrizio Cumo applied the dynamic simulation tool TRNSYS 17 (Type 56/TRNBuild) to estimate the energy consumption of the historic “Banco Napoli” building in Rome, Italy. Using a white-box approach, Cumo et al. evaluated the building’s existing energy conditions, pinpointed major issues, and then recommended suitable implementation measures [11]. An EnergyPlus model was established for the Waaghaus (weigh house), a historic building in the old town of Bolzano, northern Italy. Sensitivity analysis was then applied to determine which inputs most strongly influence the mismatch (error) between simulated and monitored indoor air temperatures [35]. To assess and validate the capability of deep learning for energy forecasting in historic buildings, Ni, Z et al. performed a case study on three public historic buildings and developed energy-consumption prediction models [10]. They evaluated eight approaches in total—seven based on deep learning—and constructed models that can output both point estimates and probabilistic forecasts, thereby addressing the multi-horizon prediction task [36]. Semerar introduced a hybrid predictive scheme that integrates Physics-Informed Neural Networks (PINNs), dynamic energy simulation, and energy-balance constraints. The framework enables one-hour-ahead prediction of indoor operative temperature and cooling electricity consumption for a historic public building in southern Italy [37].

To curb building energy demand and enhance energy efficiency, energy-consumption forecasting is a key component in the planning and optimization of building energy systems. In practical applications, prediction tasks are commonly divided into two categories: short-

term load forecasting (STLF) and long-term load forecasting (LTLF) [38]. Over the past decades, substantial research efforts have been devoted to both STLF and LTLF, leading to a wide range of forecasting approaches. Overall, existing methods can be grouped into three main classes. The first class is physics-based modeling (often referred to as “white-box” models), which relies on detailed building information and heat-balance equations. The second class includes purely data-driven approaches (“black-box” models) that learn from historical operational data using machine-learning techniques, such as support vector regression (SVR) [39], random forest (RF) [40], extreme gradient boosting (XGBoost) [41], and artificial neural networks (ANN) [42]. The third class comprises hybrid approaches (“grey-box” models), which integrate simplified physical knowledge with historical data to improve predictive performance [43,44].

Building physical energy models (white-box models), often referred to as physics-based models, describe the transient thermal response of buildings by formulating heat and mass balance equations. In the corresponding heat-balance analysis, heat exchange between the building envelope and the outdoor environment is typically represented through three mechanisms: conduction, convection, and radiation [45]. To facilitate model construction and numerical solution, a range of commercial and open-source tools is available, including EnergyPlus, Dymola, TRNSYS, DOE-2, and Matlab [46].

Among these, EnergyPlus has been continuously developed since 1997 and was first released in 2001 [47]. As an open-source program, it is widely used as a robust platform for building energy simulation. EnergyPlus generally adopts a nodal formulation, where conduction transfer functions and finite-difference schemes are commonly employed; this formulation can be viewed as essentially one-dimensional. A key benefit of the nodal approach is computational efficiency: it enables simulations over long evaluation horizons of thermal performance with relatively low computation time. Thanks to its fast execution and reliable energy-use estimation, EnergyPlus has become a popular choice for analyzing energy consumption in various buildings and energy systems, especially for large time scales such as monthly and annual simulations [48]. Some reported case studies indicate that the predicted annual electricity use can be very close to measurements, with deviations on the order of about 1% (with the prediction lower than the actual value).

TRNSYS (Transient System Simulation) was developed by the Solar Energy Laboratory at the University of Wisconsin–Madison [49]. It is a modular, transient simulation environment that allows users to flexibly assemble system components for a wide range of applications, including solar-energy systems, buildings and HVAC, renewable-energy systems, fuel cells, and cogeneration. TRNSYS is widely adopted for modeling building energy systems, particularly for solar applications and heat-pump systems [50]. Dymola (Dynamic Modeling Laboratory) originated from Hilding Elmqvist’s doctoral work in 1978, which aimed to establish a structured modeling language for large-scale continuous systems. In building-energy applications, Dymola is built on the Modelica language—an acausal, object-oriented, equation-based approach for describing physical systems such as thermal, mechanical, electrical, and control processes [45].

Beyond TRNSYS and Dymola, commonly used building energy simulation tools also include IDA ICE, DOE-2, and eQUEST. IDA ICE was developed within the Department of Building Sciences in Stockholm. DOE-2 was released in the early 1980s, and eQUEST can be viewed as an advanced implementation derived from DOE-2. In earlier work, DOE-2 and eQUEST were often employed to calibrate building energy consumption models [51,52]. In general, physics-based models tend to achieve higher predictive accuracy than purely statistical approaches [53]. Pachano, J.E. reported a case study in which only one model configuration outperformed the white-box baseline, but it required approximately twice the amount of higher-resolution data—highlighting the robustness and reliability of physics-

based white-box models for energy-use prediction [54]. In this study, a white-box simulation tool (EnergyPlus) is used to generate calibrated, heat-balance-based, itemized outputs. These outputs are then treated as interpretable Pathway layer energy-channel variables for subsequent modeling and for interpreting underlying mechanisms.

Although building technologies have progressed rapidly over the past few decades, improving energy efficiency is still constrained by limited insight into the determinants of building energy use. For example, Qinglong Meng examined how weather-related variables affect building energy consumption [55]. In addition, M. Alwetaishi proposed a design-support framework that helps architects and engineers determine an appropriate window-to-wall ratio (WWR) for each façade orientation of the same building, aiming to ensure sufficient daylight while keeping energy-load dependence as low as possible [56]. Xiangfeng Li extracted various factors from architectural forms to conduct research on the factors that determine building energy consumption [57]. Laaroussi investigated the major issues and key drivers that shape occupants' energy-related behavior and associated energy-use patterns [58]. Building on a review of the literature, Shuo Chen summarized the influencing factors into three broad categories—building characteristics, equipment and technologies, and occupant behavior—covering both internal and external determinants [59]. In a broader synthesis, R. Olu-Ajayi provided a holistic survey of research examining the diverse variables that affect energy consumption in both residential and commercial buildings [60]. However, in current studies of building energy consumption prediction, the choice of influencing factors is often guided by variables commonly reported in the literature, with an insufficient analysis of the selected factors. According to the existing literature, many researchers have not rigorously screened the influencing factors to identify the most relevant ones for their research objects. The common approach is that researchers select well-known influencing factors (such as climate, envelope structure, user behavior, etc.) without first exploring the strong correlation of the selected factors with their research objects. R. Olu-Ajayi noted that, in many studies, researchers rarely provide strong justification for why particular influencing factors are chosen [60]. From a modeling standpoint, including irrelevant variables can introduce noise and degrade predictive performance [61], which may partly explain why it remains difficult to obtain consistently top-performing building energy consumption prediction models. As a related example, Siddharth et al. applied ANN and SVM to predict the energy use of a commercial building and reported a mean absolute error (MAE) of 11.2 [62].

1.2. Contributions of This Study

Although research in this area has been growing, the inherent complexity of building energy consumption makes it especially important to identify and prioritize the most influential factors [33,34]. In a large portion of the literature, factor selection is primarily carried out using statistical techniques, expert experience, or by simply reusing the commonly adopted variables reported in previous studies. This practice suggests that the rationale for choosing influencing factors is still not sufficiently well articulated [60]. However, given the differences in building envelope, system configuration, operation mode, and human behavior among different buildings, the correlation and degree of influence of the same feature may vary in different buildings [34]. Therefore, before conducting energy consumption analysis, prediction or simulation, screening the influencing factors for a specific research object should be regarded as an indispensable preliminary step. From the perspective of model construction, including factors with weak correlation may introduce noise and affect model performance [61]. Although many screening and ranking techniques have been proposed, their results may still show method dependence and be difficult to compare across samples, and there is still room for further research and improvement

in terms of practical applicability, reproducibility and interpretability. For protected and constrained historic buildings, the above problems are more prominent: due to the need for scientific assessment before renovation, and the fact that some important factors may lack feasible practical space in practice [15]. Therefore, it is necessary to establish a screening and ranking process that can enhance comparability control and support traceable mechanistic interpretation to support implementable energy-saving decisions.

Drawing on the long-term accumulation and summary of existing research, the current consensus in the research community is that building energy consumption is primarily determined by six factor groups: (1) climate, (2) the building envelope, (3) building services and energy systems (Services and energy systems), (4) building operation and maintenance (O&M), (5) occupant activities and behavior (Occupant behavior), and (6) indoor environmental quality (IEQ) [63]. Notably, the last three categories—those closely tied to human-related influences—can affect energy use to a similar extent as, or even more than, the first three categories [64]. Based on the above classification framework, and to address the shortcomings of existing research in comparability control and mechanistic traceability explanation, this paper proposes a Driver–Pathway two-layer variable system under the condition of comparing similar buildings. The Driver layer represents the controllable operational drivers, and Pathway layer is constructed from calibrated EnergyPlus heat-balance itemized outputs to enable physically grounded explanations. For the two-layer features, a feature importance evaluation approach is adopted and integrated to rank variables, thereby identifying stable key factors and dominant heat gain/loss pathways. For validation, we apply block-wise training on the training set and a rolling-window evaluation on the test set; across rolling windows, the stability of importance distributions is examined and cross-period generalization is quantified using RMSE, CVRMSE, and R^2 . The results enhance the operability of energy-saving decisions by identifying the contribution of adjustable strategies to EUI, locking the dominant heat gain/loss pathways, and informing the prioritization of intervention strategies.

2. Methods

2.1. Feature Selection

Feature selection plays a critical role in developing building energy models, as it aims to identify the most suitable features—or feature subsets—for subsequent prediction tasks. When features are chosen appropriately, the resulting model can be simpler, trained more efficiently, and deliver improved predictive accuracy; in addition, better feature choices often enhance model interpretability. Feature selection techniques are commonly grouped into supervised, semi-supervised, and unsupervised families according to whether label information is available [65]. Since labels are available in this study, we adopt a supervised feature selection strategy. From the perspective of implementation, feature selection approaches can be categorized into filter, wrapper, and embedded methods [66]. Accordingly, this work applies these three representative approaches and compares their performance. Moreover, feature evaluation can be carried out in either a model-agnostic or a model-specific manner: model-agnostic evaluations yield results that do not depend on the chosen learning algorithm, whereas model-specific evaluations produce outcomes that vary with the selected algorithm.

The filter approach is a statistics-driven feature selection technique that relies on quantitative criteria to assess how strongly each candidate input feature is associated with the target variable (i.e., the EUI of the case buildings), and then retains the most informative features [33]. Because feature relevance is evaluated without involving any specific learning algorithm, filter methods are model-agnostic. They are also computationally efficient and typically require relatively little time to run. Depending on the evaluation scope, filter

methods can be applied to individual features (univariate filters) or to feature subsets (multivariate filters). A representative univariate filter, for example, examines the correlation between each feature and the label [67]. For this purpose, the Spearman rank correlation coefficient ρ was used because it is a nonparametric measure that quantifies the strength and direction of a monotonic association between two variables without assuming linearity or normality. Spearman's ρ is calculated on ranked values and is therefore less sensitive to outliers and scale differences than Pearson correlation. Formally, Spearman's ρ is defined as the Pearson correlation between the ranked variables $R(X)$ and $R(Y)$:

$$\rho = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}. \quad (1)$$

In the presence of ties, average ranks are used.

For continuous and ordinal features, ρ was calculated directly between each feature and EUI. For nominal categorical variables, appropriate encoding (e.g., one-hot) can be applied and ρ computed for each indicator. The Spearman coefficient ranges from -1 to 1 . A larger absolute value of ρ indicates a stronger monotonic relevance between the input feature and EUI; values close to zero indicate little to no monotonic association, while the sign of ρ indicates whether the relationship is increasing or decreasing.

Unlike filter methods, wrapper methods perform feature selection in conjunction with a machine learning algorithm, searching for the feature subset that allows the model to achieve the best predictive performance. As a result, the selected features (and their assessed importance) are model-dependent, i.e., influenced by the chosen learning algorithm. In wrapper-based selection, candidate subsets are generated according to a specified search strategy, and model training/evaluation is repeated for each subset, which typically leads to substantially higher computation time than filter methods. A commonly used wrapper-style criterion is permutation feature importance, which estimates a feature's contribution by measuring the performance drop after permuting that feature. In this study, recursive feature elimination (RFE) was employed in conjunction with an Extreme Gradient Boosting (XGBoost) regressor as the base estimator (RFE+XGBoost). RFE is a backward elimination procedure that iteratively removes the least informative feature according to the importance scores provided by the underlying model. Specifically, the XGBoost model was initially trained using the entire set of candidate features. Subsequently, feature importance was computed from the trained model (e.g., based on split gain/importance scores), and the feature with the lowest importance was eliminated. The model was refit after each elimination step, and the procedure continued until a complete ranking of all predictors was obtained (i.e., one feature remained), yielding layer-wise ranked lists of candidate features for EUI prediction (Driver layer: 30 features; Pathway layer: 20 features). XGBoost is a gradient-boosted decision tree model that learns an additive ensemble of trees, $\hat{y}_i = \sum_{k=1}^K f_k(x_i)$, by minimizing an objective composed of a data-fitting loss and a regularization term:

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (2)$$

$$\Omega(f_k) = \gamma T_k + \frac{1}{2} \lambda \|w_k\|_2^2, \quad (3)$$

where n is the number of samples, K is the number of trees, $l(\cdot)$ is the loss function (with y_i denoting the measured EUI and $\hat{y}_i$ the predicted EUI), and $\Omega(\cdot)$ penalizes model complexity (e.g., number of leaves T_k and leaf weight vector w_k) to mitigate overfitting.

Embedded methods perform feature selection inherently during model training; therefore, the selection procedure is integrated into (and typically specific to) the chosen learning

algorithm. Embedded methods are typically more computationally efficient than wrapper methods, resulting in shorter runtimes [68]. In addition, they generally incur lower computational cost and tend to be less susceptible to overfitting compared with wrapper-based approaches. The Random Forest (RF) regressor is an embedded feature selection approach, because feature relevance is assessed inherently during model training. Given a dataset (y, X) , where y denotes building EUI and $X \in R^{n \times p}$ contains p candidate features, and T is the number of trees. RF builds an ensemble of T regression trees using bootstrap samples of the training data. At each split node, RF searches only a randomly selected subset of features and chooses the split that maximizes the decrease in node impurity (for regression, typically variance or mean squared error). As a result, features that yield larger impurity reductions tend to be selected more frequently across trees, leading to implicit feature selection.

After training, feature importance is quantified by the mean decrease in impurity:

$$\text{Imp}(j) = \frac{1}{T} \sum_{m=1}^T \sum_{t \in T_m: v(t)=j} p(t) \Delta I(t), \quad (4)$$

where m indexes the trees, T_m denotes the set of split nodes in tree m , and $v(t)$ denotes the feature used for splitting at node t ; $p(t)$ is the proportion of samples reaching node t , and $\Delta I(t)$ is the impurity decrease contributed by that split. A larger $\text{Imp}(j)$ indicates a higher contribution of feature j to EUI prediction, whereas features with near-zero importance can be considered weak contributors and thus candidates for removal in subsequent screening.

2.2. Final Feature Ranking via Reciprocal Rank Fusion (RRF)

To obtain a more robust final feature importance ranking, this study aggregated three ranking lists produced by a filter method (Spearman correlation), a wrapper method (RFE+XGBoost), and an embedded method (Random Forest feature importance) using reciprocal rank fusion (RRF). Since the importance scores generated by different methods are not directly comparable in scale and interpretation, RRF operates solely on ranks, thereby avoiding bias from uncalibrated scores [69]. Specifically, the output of each method was converted into a rank $r_s(f)$ for each feature f , where $r_s(f) = 1$ denotes the most important feature. For Spearman, features were ranked by descending $|\rho|$; for Random Forest, by descending feature importance; and for RFE+XGBoost, by the elimination order (features eliminated later receive higher ranks). The fused RRF score was then computed as:

$$\text{RRF}(f) = \sum_{s=1}^M \frac{1}{k + r_s(f)} \quad (5)$$

where M is the number of ranking lists (here $M = 3$) and k is a constant that mitigates the effect of outlier high ranks from any single method. Features were finally sorted by descending $\text{RRF}(f)$ to obtain the overall ranking. The original RRF study used $k = 60$ and reported that performance was not highly sensitive to this choice; this study follows the same setting [69].

2.3. Model Interpretation by SHAP Method

An effective energy model should be both accurate and interpretable [33]. Tree-based machine-learning algorithms (e.g., XGBoost and Random Forest) can enhance interpretability by providing global feature-importance rankings; however, these scores do not quantify, for each individual sample, how much each feature increases or decreases the predicted EUI, nor do they directly indicate whether the feature effect is positive or negative. There-

fore, SHAP was adopted to interpret the developed EUI models by providing both local (sample-level) and global explanations.

The SHAP method is a game-theoretic framework for interpreting model predictions. It treats each feature as a “player” in a cooperative game, where the model’s prediction corresponds to the game outcome, and assigns each feature a contribution value accordingly [70]. SHAP values quantify each feature’s marginal contribution to the prediction value of a single sample relative to a baseline (typically the expected model output, often close to the global mean). The model exhibits additivity: the SHAP values across all features sum to the gap between the model’s prediction and the baseline (expected) output. Symmetry ensures that feature contributions depend solely on the features themselves and are independent of their order in the model; it supports both local interpretability for individual sample analysis and global interpretability for the assessment of overall feature importance [71]. SHAP is model-agnostic and can be used with any machine-learning model. To estimate SHAP values, an interpretable linear model is fitted as an explainable approximation of the original model, expressed as:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i, \quad (6)$$

where M denotes the number of input features, ϕ_i represents the SHAP value for the i -th feature, and ϕ_0 is the baseline value corresponding to the expected model output. The SHAP value quantifies the contribution of each feature (i.e., ϕ_i) by comparing the model’s predictive outcomes when the feature is included versus excluded from the model. For the calculation of feature attribution values, the SHAP value is derived in conjunction with the conditional expectation, defined as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)], \quad (7)$$

ϕ_i denotes the SHAP value for feature i ; F represents the set of all features; S denotes the feature subset (excluding feature i); $f(S)$ indicates the model prediction value using subset S ; and $|F|$ and $|S|$ denote the number of features in sets F and S , respectively [72].

2.4. CPSS Stability Selection for Top-K Feature Truncation

To assess the reproducibility of the selected Top-K feature set under different truncation levels, we adopt a standard stability selection procedure (subsampling + selection probability). Specifically, we repeatedly generate time-structure-preserving subsamples, re-run the predefined feature ranking and fusion (RRF) pipeline on each subsample, and quantify how frequently each feature appears in the resulting Top-K set. This yields feature-wise selection probabilities and a K -level stability score.

We repeat subsampling B times. Let the b -th subsample be denoted by $D^{(b)}$, for $b = 1, \dots, B$. To respect temporal dependence, subsamples are generated without shuffling time order (e.g., by drawing contiguous time blocks, or equivalently by resampling a subset of rolling windows/time segments), so that the time-series structure is preserved.

For each candidate $K \in \mathcal{K}$ (a set of candidate truncation levels), the ranking and fusion (RRF) procedure is applied to each subsample $D^{(b)}$ to obtain the Top-K feature set:

$$\mathcal{F}_K^{(b)} = \text{TopK}(\text{FuseRank}(D^{(b)})), \quad K \in \mathcal{K}.$$

For any feature f , its selection probability under Top-K is estimated as:

$$\hat{\pi}_K(f) = \frac{1}{B} \sum_{b=1}^B \mathbb{I}[f \in \mathcal{F}_K^{(b)}], \quad (8)$$

where $\mathbb{I}[\cdot]$ is the indicator function. Intuitively, a larger $\hat{\pi}_K(f)$ indicates that feature f is more consistently selected into the Top- K set across repeated subsampling.

Let $\mathcal{F}_K^{\text{full}}$ denote the Top- K feature set obtained from the full dataset. To summarize the overall reproducibility of the Top- K set, we define the (mean-type) stability score:

$$\text{Stab}(K) = \frac{1}{|\mathcal{F}_K^{\text{full}}|} \sum_{f \in \mathcal{F}_K^{\text{full}}} \hat{\pi}_K(f).$$

In addition, to reflect a conservative lower bound for the “least stable member” within the Top- K set, we also report:

$$\text{Stab}_{\min}(K) = \min_{f \in \mathcal{F}_K^{\text{full}}} \hat{\pi}_K(f). \quad (9)$$

(Optional) For interpretability-focused reporting, a threshold π_0 (e.g., 0.8 or 0.9) can be used to define “high-stability” features satisfying $\hat{\pi}_K(f) \geq \pi_0$, which supports subsequent mechanistic interpretation and reproducibility discussion.

For interpretability-focused reporting, we define the fraction of high-stability features within the full-data Top- K set as:

$$\text{Frac}(K) = \frac{1}{|\mathcal{F}_K^{\text{full}}|} \sum_{f \in \mathcal{F}_K^{\text{full}}} \mathbb{I}[\hat{\pi}_K(f) \geq \pi_0]. \quad (10)$$

where $\mathbb{I}[\cdot]$ is the indicator function and π_0 is a user-specified stability threshold (e.g., 0.8 or 0.9).

3. Case Study

3.1. Data Description

The EU project EFFESUS (2012–2016) provided a clear definition of historic buildings, which are structures built before 1945 using craftsmanship and pre-industrial techniques [73]. The European legislative and regulatory framework stipulates [74] that only buildings with a high degree of physical integrity and recognized historical and cultural features are strictly regarded as historic buildings. Based on these authoritative definitions, this paper adopts the operational definition of “buildings constructed before 1945, maintaining a high degree of physical integrity and possessing recognized historical and cultural characteristics” for historic buildings, in order to clearly define the research subjects and statistical scope.

Based on the above definitions and criteria, case selection in this study followed three principles: object comparability, practical conservation demand, and climate representativeness. Accordingly, four in-use historic buildings were selected as the study objects from the officially published List of Historic Buildings in Shenyang City resulting from the city-wide survey [75]. Shenyang was chosen as the case city for three main reasons.

First, Shenyang provides a relatively complete urban context with an explicit policy orientation toward conservation through use. Shenyang is designated by the State Council as a National Historic and Cultural City (the second batch) and has over 2300 years of urban development history, with a large stock of historic buildings, which offers a more comprehensive urban background for research on similar historic building typologies [76]. More importantly, relevant municipal policy documents explicitly state the principle of “promoting conservation through utilization” and call for “actively studying measures for the conservation and utilization of historic buildings” [77], ensuring that this study is aligned with real management needs and has stronger practical applicability. Second, Shenyang is located in China’s severe cold region [78]. It has pronounced winter heat-

ing demand, while field measurement studies have also demonstrated a risk of summer overheating [79]. This enables the Shenyang cases to reflect HVAC-related energy characteristics in both winter and summer, making the study more targeted and representative for operational energy identification and energy-saving decision-making in historic buildings.

Third, Shenyang has established a relatively systematic conservation foundation for historic buildings. Since the first city-wide historic building survey in 2017, a total of 236 historic buildings have been officially announced by the time this paper was completed, and the city has continuously promoted plaque installation, the “Readable City” initiative, digital archiving, and the preparation of conservation guidelines/drawing documents [80]. This indicates not only a large historic building stock but also a relatively mature management system, which facilitates access to continuous, standardized, and verifiable objects and background documentation for research.

Among these 236 historic buildings, 162 meet the EFFESUS (2012–2016) definition of historic buildings. Excluding 30 buildings that require repair before being put into use, and 11 other objects such as ancient wells and water towers, the remaining sample consists of 121 buildings. Following commonly used classification approaches in the academic literature [73,74], these 121 historic buildings were grouped by building use into four categories, (A) residential, (B) business and commercial service, (C) public management and public service, and (D) industrial, accounting for 7%, 56%, 30%, and 7%, respectively (see Figure 1).

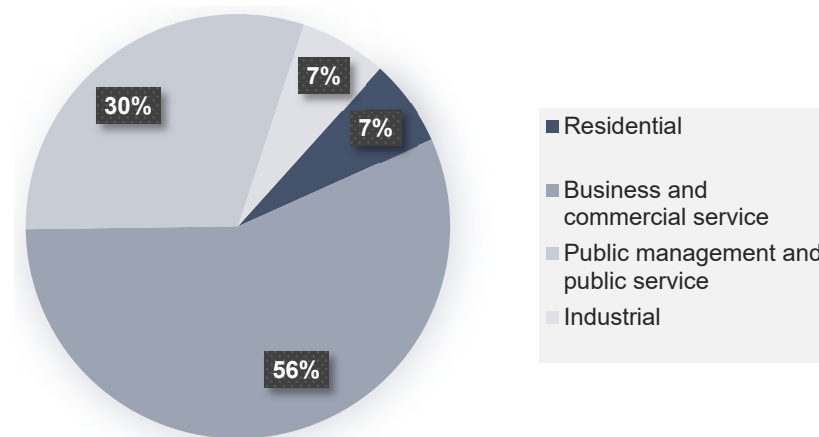


Figure 1. Composition of historic buildings by use category.

Based on this distribution, and to ensure representativeness of the study sample, we selected Category B (business and commercial service), which accounts for the largest proportion of the historic buildings, as the focus of the subsequent analysis. Considering the business-type characteristics of Category (B) historic buildings in Shenyang, and from the perspective of building operational energy research, Category (B) buildings were further subdivided into six use types: B1 (retail and service), B2 (snack bar and cafe), B3 (hotel), B4 (restaurant), B5 (office), and B6 (cultural park) (see Figure 1). Field survey results indicate that retail and service shops are often distributed in adjacent clusters; therefore, they were grouped as B1. B2 was classified using “low cooking demand and typically no dedicated kitchen space” as the main criterion. Some B1 and B2 buildings exhibit mixed-use patterns, where the ground floor is commercial while the second/third floors are residential or accommodation. In this study, buildings were categorized primarily by the ground-floor business type; however, in the energy simulation, parameters were input separately for different floors to avoid interference from functional mixing. Therefore, the categorization itself does not affect the interpretation of simulation results.

Among B3–B6, B3 (hotel) shows large variations in sample scale and complex management structures; moreover, key operational information such as occupancy rate is business-sensitive, resulting in generally low cooperation and making it difficult to establish a verifiable data basis. B4 (restaurant) typically includes an independent back-of-house kitchen with intensive cooking, requiring dedicated grease exhaust and make-up air systems; internal loads and ventilation/exhaust systems differ substantially, leading to weak cross-sample comparability. B5 (office) has a relatively small number of samples. B6 (cultural park) is mostly a composite commercial area formed by a combination of historic buildings and new buildings; boundary conditions are complex and the influence of the new-built part is difficult to isolate. Considering both proportion and comparability, B1 and B2 together account for approximately 65% (see Figure 2). In addition, they involve diverse business types and high operational intensity, implying more prominent risks of use-related damage and a stronger need for operational energy management. Therefore, B1 and B2 were selected as the sampling scope in this study.

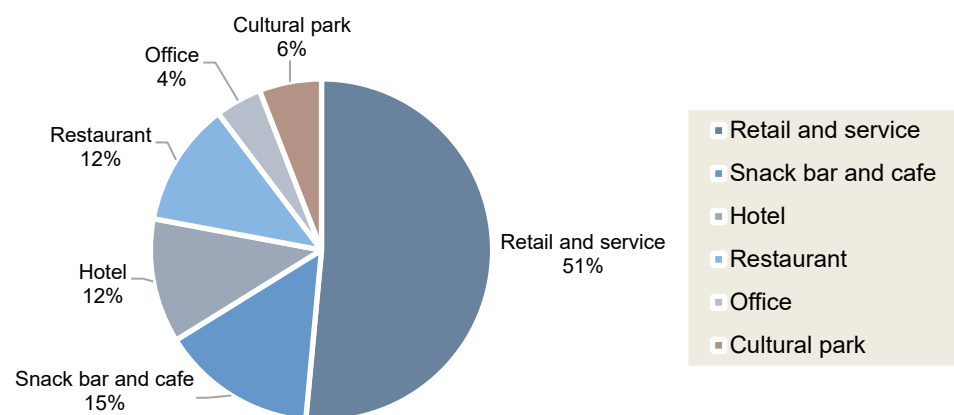


Figure 2. Proportional breakdown of commercial-service historic buildings (B1–B6).

Within the selected scope, building form can be mainly divided into two types: standalone buildings and row (terraced) buildings, accounting for approximately 51% and 49%, respectively. Most row (terraced) buildings are attached to non-historic buildings; within the row-building samples, the proportion that is not connected to adjacent new constructions is about 40%. Considering that the technologies and operational conditions of the new-built parts may introduce confounding effects and reduce the accuracy of identifying energy-use characteristics attributable to the historic buildings themselves, this study further used standalone buildings as the screening target.

This study selected four historic buildings still in use in Shenyang City, China, as the study objects. These buildings were all constructed in the 1920s, approximately a hundred years ago; and have been included in the “List of Historic Buildings in Shenyang City”. Unlike ordinary buildings, historic buildings often have missing or incomplete archival materials due to their long history, and the existing literature does not fully meet the research needs. Therefore, this study first carried out on-site surveying and mapping and drew the floor plans and elevations of the buildings. Then, based on the surveying and mapping results, a three-dimensional geometric model was established to acquire the geometric parameters and spatial form information of the buildings. During the surveying and mapping process, on-site investigations were conducted simultaneously to supplement and collect information on the building envelope, usage functions, and operational characteristics, etc., to form a relatively complete basic data set. Finally, the model was built and the calculation process was organized through the OpenStudio Application v1.5.0, and the building energy consumption was simulated using EnergyPlus (v22.2.0). The discussion in

this article on the need and application of energy simulation does not imply the implementation of real energy optimization processes. Instead, energy simulation is regarded as an auxiliary decision-making tool serving specific research scenarios. Its application scope is limited to historical building objects that have a high degree of relevance to the case in this article in terms of functional attributes, usage conditions, protection constraints, and environmental background. Within this framework, the value of energy simulation does not lie in generating universally applicable energy consumption optimization conclusions, but rather in providing more evidence-based judgments from the perspectives of protection priority, continued applicability, and energy conservation. On this basis, this paper aims to explain the mechanism and operational driving factors behind energy consumption disparities among similar historic buildings, instead of performing statistical extrapolation at the stock level. Considering the scarcity of samples of century-old historic buildings and the necessity to meet screening criteria such as being in use within the same city and having consistent control conditions, four comparable cases were finally chosen for comparison. By controlling conditions and comparability constraints (with differences restricted within a narrow range), the confounding effect is reduced, and the interpretability and engineering applicability of the conclusions are improved.

Given that historic buildings are subject to legal protection requirements and conservation technical principles (such as the principle of minimum intervention), key elements such as their envelope structures and spatial layouts often lack feasible renovation space. Unlike new buildings or those that can be freely renovated, even if certain influencing factors have high importance scores in the feature importance analysis, they may be difficult to transform into executable energy-saving renovation strategies due to conservation constraints. To enhance the applicability of research conclusions and avoid important but ineffective conclusions, this paper attempts to control the differences in key attributes among samples as much as possible during the selection of research objects, reducing sample heterogeneity and potential confounding effects. This is to prevent the artificial inflation of the importance of certain factors due to excessive individual differences, which in turn limits the generalizability of the conclusions and makes it difficult to guide actual conservation and energy-saving renovation work. Based on this, this paper first conducts comparability control and feature stratification definition of the research object: the influencing factors are divided into protected and constrained structural characteristics (non-intervenable factors) and factors that can be improved through operation management or equipment updates (intervenable factors). Among them, the non-intervenable factors are further classified into control conditions (completely consistent in all cases) and comparability constraints (differences exist among cases but are controlled within a narrow range). Control conditions are used to define the research boundaries and reduce confounding, and therefore are not included in the subsequent model input; the summary of control conditions is shown in Table 1.

To obtain a more reliable ranking of influencing features, this paper further imposes comparability constraints on several key geometric and structural indicators. Specifically, it prioritizes the selection of historic buildings with well-controlled differences in scale and form as research objects. The corresponding sample screening criteria and control conditions are summarized in Table 1. This approach aims to minimize the interference of scale effects and form differences on the ranking conclusions. Taking the gross floor area (GFA) as an example, if the samples have order-of-magnitude differences in GFA (such as 150–2000 m²), when the total energy consumption is set as the prediction target, the model may prioritize the use of scale variables to explain the energy consumption differences; meanwhile, for tree-based learners (such as Random Forest, XGBoost), continuous variables may be overestimated due to more potential split points under feature importance measures

based on split improvement, thereby masking the influence of other factors. Regarding highly redundant variables, such as when height and S/V ratio have obvious overlapping representations, this paper makes choices based on the principle of avoiding repetitive representations: their differences have been controlled through sample screening, and their main influences can be indirectly reflected by comprehensive indicators such as S/V ratio, so they are not used as input variables for the model.

Table 1. Summary of control conditions for sample screening.

No.	Influence Factor	Control Specification	Data Source	Applied in Screening	Notes
1	Year of construction	1920–1925	Historic files	✓	Century-old historic building
2	Urban location	Central business district (CBD), Shenyang	Field survey	✓	Comparable pedestrian flow within the CBD
3	Street context	Street-facing retail frontage	Field survey	✓	Comparable retail activity; unobstructed facade exposure (no major adjacent shading)
4	Energy carrier	Electricity-only	Field survey	✓	Plug-load/equipment end uses are excluded due to high tenant heterogeneity
5	Building form	Detached Rectangular plan Two stories Flat roof No courtyard	Historic files Field survey	✓	Geometry controlled (plan/height/roof type) to reduce S/V ratio-related variability

Note: ✓ indicates that the criterion was applied in screening.

To compare the differences in the ranking of factors influencing building energy consumption due to different structural types and usage functions, the research subjects were divided into two groups according to structural type: brick–wood structure ($n = 2$) and brick–concrete structure ($n = 2$); at the same time, they were also divided into two groups based on usage function: mixed commercial–residential function ($n = 2$) and pure commercial function ($n = 2$). Structural type and building use are categorical variables and were encoded as binary dummy variables prior to model training: brick–wood structure = 1 and brick–concrete = 0; mixed-use = 1 and purely commercial = 0. This coding scheme is equivalent to one-hot encoding for two-level categories and can be directly used as input features for subsequent analyses.

After ensuring the comparability of the research subjects, this paper constructs two sets of input feature sets and conducts sorting operations respectively. The first set is the Driver layer (driver/handle variables), which includes outdoor meteorology and solar radiation, static parameters of building geometry and construction, and a small number of operation/set inputs. This set is used to explain “which operable driver factors mainly lead to the EUI difference”. The second is the Pathway layer, which adopts the calibrated EnergyPlus heat balance itemized output (such as infiltration total heat gain/loss energy, windows total heat gain/loss energy, surface inside face conduction heat transfer energy, internal heat gain, etc.), and is used to identify “which heat channels mainly lead to the EUI difference”. The Pathway layer is constructed primarily as an interpretability layer to explain how the simulated EUI is distributed across major heat-transfer pathways, rather than as a stand-alone validation of real-world physical mechanisms. To address the concern of potential simulation-to-simulation “closed logic”, a physics-based plausibility check is further conducted for one representative pathway—infiltration heat loss—using independent in situ measurements.

3.2. Model Development

To ensure the comparability and reproducibility of time-series prediction, prevent future information leakage, and take into account computational efficiency, this study adopts a two-stage validation framework of block-based development and rolling evaluation. The data consists of hourly observations of each building for a complete natural year (excluding 29 February), totaling 8760 points for each of the four buildings. All hourly energy items (including Pathway layer energy variables and the hourly EUI target) are expressed in terms of energy per square meter of building area (kWh/m^2), obtained by normalizing with the total building area. The research site is located in Shenyang, which has typical strong seasonal climate characteristics: cold and dry in winter and hot with concentrated precipitation in summer, as reflected in the EPW weather data.

During the development stage, a time-sequential block splitting approach was adopted, with the last quarter of the year (1 October—31 December) serving as the external test set and the rest of the time as the training set. For the data in Tables 2 and 3, all preprocessing steps, such as missing value handling, encoding, and standardization, were only performed on the training set and then applied to the test set. Feature importance ranking, RRF, and hyperparameter search were all conducted within the training set through time-series cross-validation to ensure that no test information was used for feature selection. To eliminate the influence of building size on the comparability of energy-related metrics, EnergyPlus outputs (J) were converted to kWh and normalized by gross floor area (GFA) in this paper, yielding kWh/m^2 . In contrast, the Driver layer variables maintained their original physical quantities and attribute values without area normalization to preserve the original scale information of factors such as meteorological conditions, building size, and operational setpoints. It should be noted that the EnergyPlus model was calibrated and validated in this paper: the error between the Zone Mean Air Temperature generated by the model and the measured indoor air temperature was controlled within $\pm 1\text{ }^\circ\text{C}$ during the analysis period. Therefore, the simulation outputs used in the subsequent sections of this paper correspond to the calibrated operation state that can reproduce the measured indoor temperature, rather than the idealized indoor temperature based on thermal comfort target assumptions.

Table 2. Summary of Driver layer variables and data sources.

Category	Driver Layer Variable	Abbrev.	Type	Source	Area-Normalized (by GFA)	Unit
Meteorological Conditions	Outdoor air dry-bulb temperature	Outdoor dry-bulb temp	Weather input	EPW	NO	$^\circ\text{C}$
	Outdoor air relative humidity	Outdoor RH	Weather input	EPW	NO	%
	Wind speed		Weather input	EPW	NO	m/s
	Wind direction	WinDir sin/WinDir cos	Weather input	EPW	NO	deg
Solar Radiation	Diffuse solar radiation		Weather input	EPW	NO	Wh/m^2
	Direct solar radiation		Weather input	EPW	NO	Wh/m^2
Building Size & Form	Gross floor area	GFA	Model input	Field survey	NO	m^2
	Surface-to-volume ratio	S/V ratio	Derived	Field survey → EnergyPlus model input	NO	$1/\text{m}$
	Building Main Orientation	Orient sin/Orient cos	Derived	Field survey	NO	deg
	Number of Rooms	Room Count	Model input	Field survey	NO	count
Envelope Thermal Properties	Overall heat transfer coefficient of the building envelope	U-Value	Derived	Field survey → EnergyPlus model input	NO	$\text{W}/(\text{m}^2 \cdot \text{K})$
	Window-to-wall ratio	WWR	Derived	Field survey → EnergyPlus model input	NO	%
Structural type & Building use	Structural type		Categorical (encoded)	Field survey	NO	—
	Building use	Bldg Use	Categorical (encoded)	Field survey	NO	—
Infiltration	Infiltration Air Change Rate	Infil ACH	Model input	Field survey → EnergyPlus model input	NO	$1/\text{h}$

Table 2. Cont.

Category	Driver Layer Variable	Abbrev.	Type	Source	Area-Normalized (by GFA)	Unit
Operational Setpoints	Zone Thermostat Heating Setpoint Temperature	ZT Htg Setp Temp	Model input	EnergyPlus model input	NO	°C
	Zone Thermostat Cooling Setpoint Temperature	ZT Clg Setp Temp	Model input	EnergyPlus model input	NO	°C
Occupancy	Occupant count	Occ Count	Model input	Field survey → EnergyPlus model input	NO	persons

Notes: → indicates that field survey data were used as inputs to the EnergyPlus model.

Table 3. Summary of Pathway layer variables and data sources.

Category	Pathway Layer Variable	Abbrev.	Type	Source	Area-Normalized (by GFA)	Unit
Envelope Conduction	WALL: Surface Inside Face Conduction Heat Transfer Energy	Wall Cond Heat Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	ROOF: Surface Inside Face Conduction Heat Transfer Energy	Roof Cond Heat Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	FLOOR: Surface Inside Face Conduction Heat Transfer Energy	Floor Cond Heat Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
Windows	Windows Total Heat Gain Energy	Win Heat Gain Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	Windows Total Heat Loss Energy	Win Heat Loss Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	Windows Total Transmitted Solar Radiation Energy	Win Trans Solar Rad Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
Occupancy & Lighting	People Total Heating Energy	Occ Htg Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	Lights Total Heating Energy	Ltg Htg Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
Infiltration	Infiltration Total Heat Gain Energy	Infil Heat Gain Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²
	Infiltration Total Heat Loss Energy	Infil Heat Loss Energy	Simulation output	EnergyPlus output	Yes	kWh/m ²

In the reporting stage, after finalizing the feature set and model, a rolling evaluation was adopted with an expanding training window starting from an initial 30-day period, followed by a 7-day step forward. At each step, the model was trained with the accumulated historic data and evaluated on the subsequent 7-day test window (30-day initial training, 7-day test, 7-day step). The generalization performance and stability were evaluated across different windows throughout the year, and RMSE, CVRMSE, and R^2 were used to report the overall level, fluctuation range, and the worst window performance. The purpose of using rolling was not to make the model more accurate. The value of rolling did not lie in improving the model's precision but in making the estimates of RMSE, CVRMSE, and R^2 closer to the future performance in practical applications. Specifically, training on past windows and predicting the immediately following future windows avoided the information leakage of adjacent samples introduced by random splitting in the strongly time-correlated hourly EUI (kWh/m²) data, which could lead to inflated metrics. Secondly, the repeated evaluation of rolling across different windows throughout the year could obtain the performance distribution rather than a single value, making it more reliable and suitable for the research setting of this paper to characterize stability and identify failure risks under seasonal shifts and extreme weather conditions.

Among the three feature selection methods adopted in this paper, namely filter, wrapper, and embedded, the filter method ranks features based on their statistical association with the target variable and does not involve the training of the predictor or the prediction output on the test set. Therefore, in this paper, the Spearman rank correlation is used to obtain the correlation ranking of candidate features as a model-independent baseline

result. RMSE, CVRMSE, and R^2 are generalization performance metrics of the prediction model, but the Spearman filter ranking does not generate prediction outputs. It is a univariate correlation ranking, so this paper does not report RMSE, CVRMSE, and R^2 for the Spearman method. This ranking is only used for comparison and discussion with the subsequent wrapper (XGBoost-RFE) and embedded (Random Forest) ranking results based on prediction models. The rolling evaluation metrics are only used to assess the wrapper (XGBoost-RFE) and embedded (Random Forest) methods based on prediction models.

In the model validation stage, to evaluate the prediction performance of the model on the test set, this paper adopts three commonly used statistical indicators: root mean square error (RMSE), the coefficient of variation root mean square error (CVRMSE), and the coefficient of determination (R^2). R^2 is reported as a percentage in this paper. Among them, RMSE is used to measure the overall deviation between the predicted values and the measured values; CVRMSE normalizes the RMSE based on the mean of the measured values, facilitating the comparison of errors between data of different magnitudes; R^2 is used to represent the degree to which the model explains the fluctuations (variance) of the measured data, that is, the goodness of fit. The definitions of each indicator are as follows:

Let the true value of the i -th hour within the test window be $y_{\text{data},i}$ (normalized according to the total floor area GFA of the building), the predicted value by the model be $y_{\text{predict},i}$. Here, n is the number of hours included in this test window, and $\bar{y}_{\text{data}}$ is the mean of the true values of this window. Then:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{\text{predict},i} - y_{\text{data},i})^2} \quad (11)$$

$$\text{CVRMSE}(\%) = \frac{\text{RMSE}}{\bar{y}_{\text{data}}} \times 100 \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

The three indicators are independently calculated within each rolling window. The final report includes the average RMSE, CVRMSE and R^2 values for the entire year window, as well as the fluctuation range and the worst performance of the window. Generally speaking, the smaller the CVRMSE and RMSE and the larger the R^2 , the more accurate the prediction. This paper uses CVRMSE as the main comparison indicator. When the difference in CVRMSE is not significant, RMSE and R^2 are also combined to assist in the judgment.

Since CVRMSE is normalized by the mean value $\bar{y}_{\text{data}}$ of the test segment ($\text{RMSE} / \bar{y}_{\text{data}} \times 100$), its cross-period comparison needs to be interpreted together with RMSE and R^2 .

3.3. Physics-Based Plausibility Check

The plausibility check is demonstrated using the case with the most complete measurement coverage and representative operation (e.g., Building 2) and is restricted to F1 (retail) because its operating schedule provides clearer unoccupied periods, enabling more reliable identification of CO₂ decay segments with minimal internal generation and thus more robust ACH inference; in contrast, F2 (residential) exhibits less regular occupancy and more frequent disturbances (e.g., intermittent activity or window opening) that weaken the “no-source” assumption. Indoor CO₂/temperature/RH are measured using an HZ-9110A monitor (Dongguan, China) (one record per minute, timestamped at:37 s within each

minute); to ensure a consistent time basis with EnergyPlus outputs, indoor timestamps are shifted by -37 s and floored to exact minutes (INDOOR_SHIFT_SECONDS = 37) before hourly aggregation. During selected no-source periods, the decay of indoor CO₂ above an outdoor background C_{out} is approximated by the mass balance

$$\frac{d(C - C_{out})}{dt} = -ACH(C - C_{out}), \ln(C - C_{out}) = -ACH \cdot t + b$$

where the negative slope provides the effective air-change rate $ACH(h^{-1})$. Using the inferred ACH , the infiltration sensible heat-loss rate is estimated as

$$\dot{Q}_{inf,sens}(t) = 0.33 \cdot ACH(t) \cdot V \cdot \max(T_{in}(t) - T_{out}(t), 0)$$

where V is the zone volume (m^3), T_{in} and T_{out} are indoor and outdoor temperatures ($^{\circ}C$), and 0.33 is the engineering approximation of $\rho c_p / 3600$. Integrating $\dot{Q}$ over the monitored week yields the estimated infiltration sensible heat-loss energy $E_{inf,sens}$ (kWh).

To ensure a consistent metric, the CO₂ – ΔT first-order estimate is compared with the EnergyPlus sensible infiltration heat-loss output (Zone Infiltration Sensible Heat Loss Energy) at the same aggregation level (hourly/weekly), supporting a reasonable order of magnitude and trend for the infiltration-loss pathway. This check is used solely to support physical interpretation of the pathway variable and does not affect feature ranking or model training.

As summarized in Table 4 for the heating week (18–24 December 2023), the CO₂-based estimate is 277.79 kWh versus 364.48 kWh from EnergyPlus (weekly bias -23.79%), supporting a reasonable order of magnitude for the infiltration-loss pathway. This check is used solely to support physical interpretation of the pathway variable and does not affect feature ranking or model training.

Table 4. Winter-week infiltration sensible heat-loss plausibility check summary (Floor 1).

Metric	EnergyPlus (EP)	CO ₂ -Based Estimate	Notes
Zone	F1	F1	Comparison is for the same zone.
Zone volume (m^3)	452.61	452.61	Used to convert ACH to volumetric flow and sensible heat loss.
Floor area (m^2)	125.72	125.72	Used for area-normalized reporting ($kWh/m^2 \cdot week$).
Number of hourly samples	168	168	$7 \times 24 = 168$ h in the winter week.
Weekly sensible infiltration loss (kWh)	364.48	277.79	EP: Zone Infiltration Sensible Heat Loss Energy, aggregated over the week.
Normalized ($kWh/m^2 \cdot week$)	2.90	2.21	Facilitates cross-building/zone comparison.
Weekly bias (%)	—	-23.79	Defined as $(\text{Estimate} - EP) / EP$; negative means the estimate is lower than EP.
Hourly CVRMSE (%)	—	47.43	Large hour-to-hour variability is common (door-driven exchange, short-term mixing, etc.).
Hourly NMBE (%)	—	23.93	As defined in the script (EP as reference); positive indicates EP is higher than the estimate on average.
Valid CO ₂ decay windows	—	114	Number of decay windows used to infer effective ACH.
Median ACH (1/h)	—	0.524	Median effective air-change rate inferred from decay fits.

Note: Found 114 valid decay windows; using interpolated ACH (median = 0.524 1/h). Time range (raw timestamps): 2023-12-18 00:00:37–2023-12-24 23:59:37; time range (minute-aligned for matching with EP): 2023-12-18 00:00:00–2023-12-24 23:59:00.

4. Results

4.1. Feature Selection Process

In this study, the term “feature” is defined as the input feature (predictor variable) of the model; this section will report and discuss the feature importance ranking. To ensure a fair comparison among the three feature-selection approaches (filter, wrapper, and embedded methods), the feature ranking was first obtained on the same training dataset. All methods utilized the same candidate feature set and preprocessing procedures, so that the differences in the rankings could reflect the characteristics of each approach.

Figure 3 reports the Spearman-based filter ranking on the training dataset. In the Driver layer, the Infil ACH value is the most prominent, indicating that the difference in building airtightness is almost the strongest correlation with EUI. WWR indicates that the total area of windows and their orientation represent different mechanisms and are coupled. It should be noted that Spearman reflects statistical association rather than causation (correlation $\neq$ causation). Therefore, individual features (such as Total WWR) showing a certain direction may contradict common sense symbols, but cannot be interpreted as the causal influence of this variable on energy consumption. It is more likely due to the coupling between variables. This issue will be further cross-validated and discussed in the subsequent wrapper and embedded methods. Room Count and GFA rank high, suggesting that area may bring higher system operation loads. Occ Count and Thermostat Setpoint Temperature are at medium-high levels, indicating stability in this study. The correlation of meteorological variables is weak, indicating that building physical performance and operational differences can better explain the impact on EUI than instantaneous weather differences.

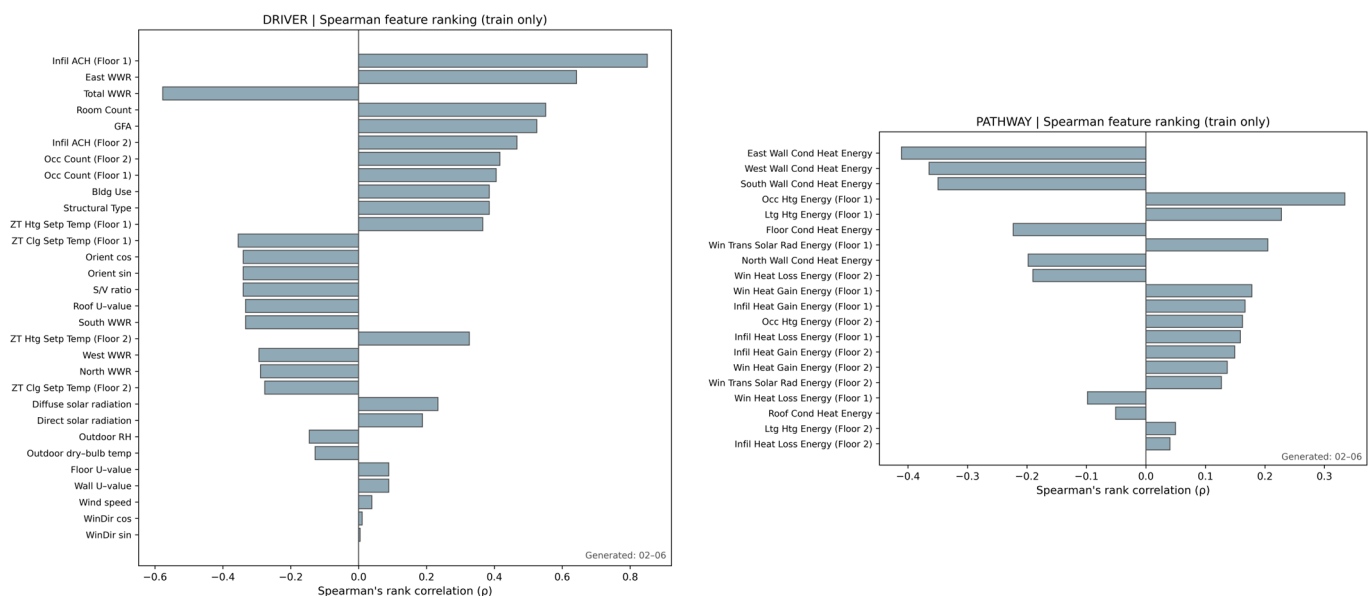


Figure 3. Spearman feature ranking for the DRIVER and PATHWAY layers (train set only).

The most prominent pathways in the Pathway layer come from the categories of Envelope Conduction and Occupancy and Lighting (hereafter referred to as Occupancy and Lighting). Wall Cond Heat Energy ranks high with $|\rho|$ values of 0.411, 0.365, 0.350, and 0.198, indicating that the heat flux variation in the envelope wall is an important channel. Floor Cond Heat Energy is also in the top ranks with $|\rho| = 0.223$, suggesting that heat transfer through the ground floor slab cannot be ignored. Occ Htg Energy and Ltg Htg Energy are positively correlated with EUI in the samples of this study, indicating that they may increase EUI by raising the cooling demand or extending the system operation time. The pathways related to Windows are in the middle to high positions, showing that the solar transmission and heat gain through windows have a stable and monotonic relationship with energy consumption. Infil Heat Gain/Loss Energy is strongly correlated with Infil ACH in the Driver layer, mutually confirming each other. It should be noted that EnergyPlus uses a directional symbol convention for the output of Envelope Conduction (the symbol convention for the inside face is consistent with Surface Inside Face Conduction Heat Transfer Energy, and the symbol for the outside face-related variables is opposite). Therefore, the positive and negative signs of the Spearman correlation coefficients in this study only represent the statistical monotonic relationship, and their physical meanings

need to be interpreted in combination with the Gain/Loss direction of the output variables. This is not the content discussed in this article.

Figure 4 reports the wrapper-based feature ranking performed using XGBoost as the base estimator through recursive feature elimination (RFE) on the training dataset. RFE is applied iteratively with XGBoost as the base estimator to rank features according to their contribution to model performance. To enhance interpretability and perform consistency checks, the plot also reports the importance based on XGBoost gain from the same estimator. It is important to note that RFE rankings arise from a wrapper-based feature selection strategy, whereas XGBoost gain reflects an intrinsic, model-embedded importance measure. Presenting both is intended to assess their agreement rather than to introduce an additional feature-selection paradigm.

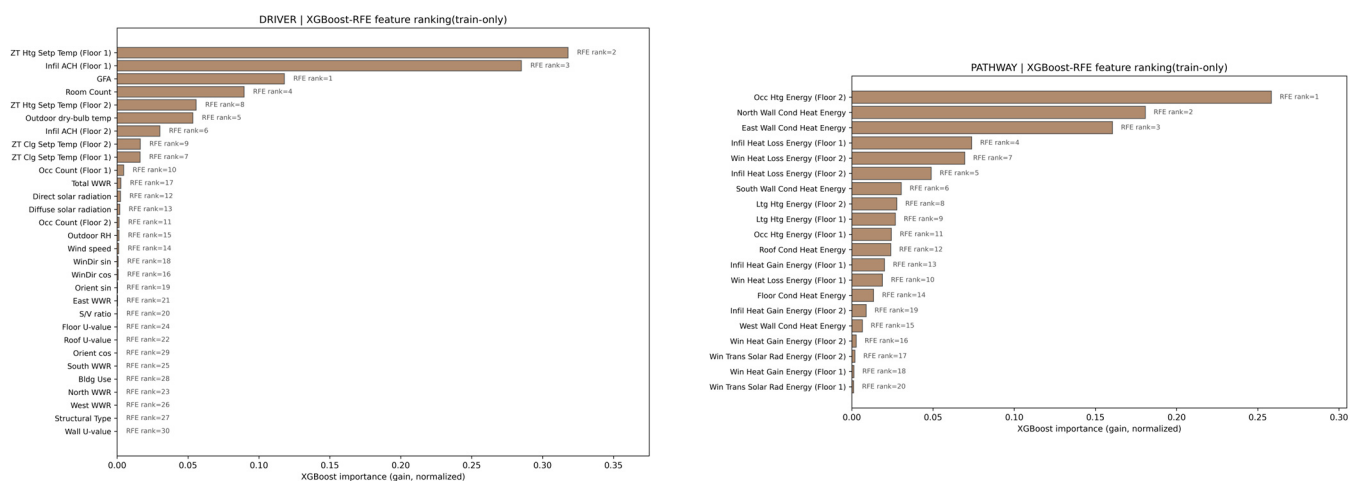


Figure 4. XGBoost-RFE feature ranking for the DRIVER and PATHWAY layers (train set only).

As shown in Figure 4, Thermostat Setpoint Temperature in the Driver layer is the primary feature, indicating that under this data and target metric, the contribution of the heating control strategy to EUI is stronger than that of the Envelope Thermal Properties. Infil ACH remains crucial, suggesting that the difference in air tightness is a stable and operable feature. Building Size and Form (hereafter referred to as Size and Form) still has explanatory power: GFA ≈ 0.118 and Room Count ≈ 0.090 rank among the top four, indicating that GFA and spatial division significantly affect the operational load. Unlike Spearman, in the multivariate prediction framework, WWR ranks much lower, which is consistent with the situation where the effect of windows may be coupled by variables such as settings, infiltration, area, and weather. Outdoor dry-bulb temp ranks sixth, suggesting that after controlling for building and operational differences, weather remains an important feature that cannot be ignored. The Pathway layer indicates that Occupancy and Lighting take the lead: Occ Htg Energy (Floor 2) ≈ 0.258 is the most important feature, and Ltg Htg Energy (Floor 1/2) and Occ Htg Energy (Floor 1) rank among the top ten, indicating that the heat release from occupants and lighting in the sample is highly correlated with EUI changes. Envelope Conduction is also important: North/East/South Wall Cond Heat Energy rank among the top seven, with North Wall ≈ 0.181 and East Wall ≈ 0.161 , reflecting that wall conduction heat transfer energy is an important load formation mechanism. Infil Heat Loss Energy and Win Heat Loss Energy rank high in this model, indicating that Heat Loss Energy has a stronger impact on EUI in this model; in contrast, Win Heat Gain Energy and Win Trans Solar Rad Energy rank lower in the model, possibly due to their stronger dependency or absorption by other strong variables.

In Figure 5, Random Forest is used as an embedded ranking method. It first fits the model on the training dataset of all candidate features and then extracts the model-derived

feature-importance ranking to quantify each feature's relative contribution to the EUI. Finally, the features are sorted from high to low and the Top-K features are selected. The Driver layer based on the RF importance ranking shows that the main influencing factors of EUI are concentrated in three types of features: Infil ACH, Meteorological Conditions, and Windows. Among them, Infil ACH (Floor 1) has a much higher importance than other features, indicating that the difference in air tightness remains the strongest feature in the RF framework. Outdoor dry-bulb temp ranks second, and meteorological quantities such as Outdoor RH, Direct solar radiation, Diffuse solar radiation, and Wind speed are also among the top-ranked features, suggesting that in the RF's nonlinear fitting, the hourly variation in EUI is more strongly associated with external meteorological conditions. Total WWR ranks third, indicating that the total window area is affected by meteorological conditions and has a key impact on the load. Among the setpoint variables, ZT Clg Setp Temp (Floor 2) ranks in the top ten, but ZT Htg Setp Temp (Floor 2) ranks lower. Parameters such as GFA, S/V ratio, U-value, and Orient sin rank lower in RF, suggesting that these features may have smaller variations within the sample or are absorbed by the top-ranked features, thus being marginalized in the importance allocation of RF.

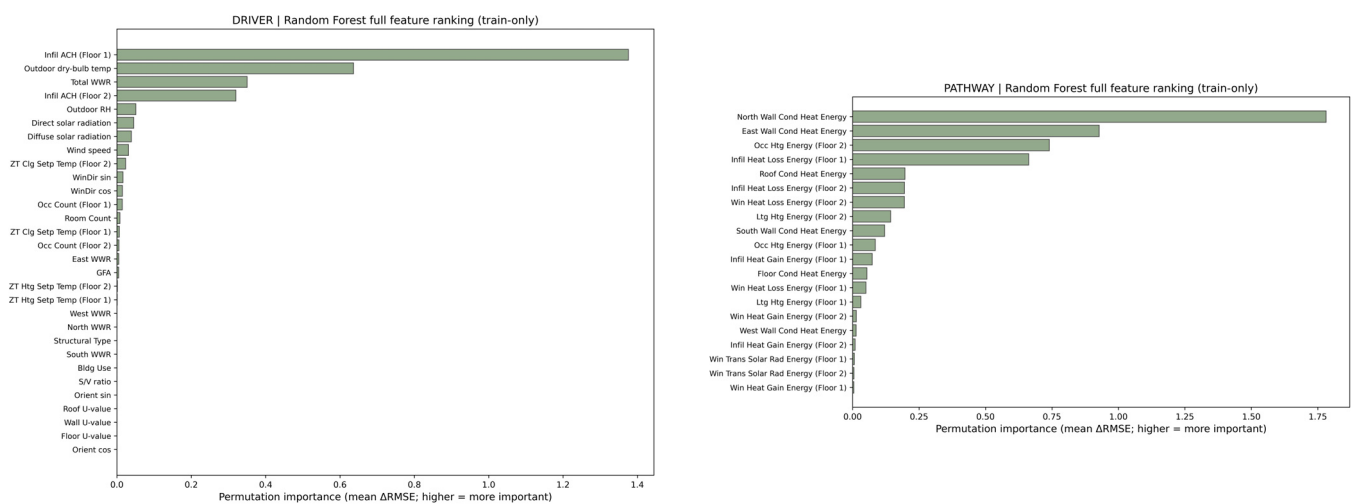


Figure 5. Random Forest feature ranking for DRIVER and PATHWAY layers (train set only).

The RF sorting in the Pathway layer is visible, where North/East Wall Cond Heat Energy rank in the top two positions with significantly higher weights than the other features, indicating that Wall Cond Heat Energy is the most critical load formation mechanism. South Wall Cond Heat Energy and Roof Cond Heat Energy are also among the top-ranked, further reinforcing the conclusion that envelope conduction is dominant. Occupancy and Lighting are equally important: Occ Htg Energy (Floor 2) ranks among the top three, and Ltg Htg Energy (Floor 2) is in the top ten, suggesting that the heat release from occupants and lighting exhibits a relatively strong influence on EUI variation across the samples. Meanwhile, Infil Heat Loss Energy (Floor 1/2) and Win Heat Loss Energy (Floor 2) are both in the top positions, indicating that the above three pathways have a stronger discriminatory power for EUI. In contrast, Win Trans Solar Rad Energy and Win Heat Gain Energy rank lower, suggesting that their effects may be season-dependent or absorbed by the top three features.

The three feature-selection approaches in this paper each have their own systematic biases: Spearman sometimes ranks a group of homologous information variables very high, unable to distinguish the true contributors. If there is a coupling situation, Spearman captures marginal associations, which cannot reflect the contribution after controlling for other variables and cannot reflect interactions. XGBoost-RFE may redistribute the

explanatory power among correlated features, and gain prefers features that cause greater loss in early splits, so variables of the same mechanism may be represented alternately in different folds. If RF uses permutation importance, it may concentrate importance on features that more strongly affect prediction error when permuted; when features are correlated, importance can be shared or shifted among correlated variables, which may lead to structural bias and instability. The purpose of choosing RRF is to fuse at the ranking level without changing the internal mechanisms of each method. It is less sensitive to the occasional overestimation or underestimation of a certain method; at the same time, it rewards stable features that are consistently ranked high by multiple methods and moderately downgrades features that are extremely high in a single method but ranked low in other methods, thereby obtaining a more stable core feature set.

4.2. Model Performanc

The time generalization performance of the feature importance models developed by XGBoost-RFE and Random Forest algorithms was evaluated based on the rolling indicator table. In the XGBoost-RFE model, the evaluation was conducted under a 47-fold rolling extended-window training, where each fold used the training set accumulated from the starting date to the current period and was tested in the subsequent 672 h period.

Table 5 presents a 47-fold rolling evaluation comparing a single-model RF-only baseline against the proposed RRF+XGB (Top-10) pipeline. Relative to the RF-only baseline, RRF+XGB improves both overall accuracy and stability in both layers. For the Driver layer, mean RMSE decreases from 0.156 to 0.135 (−13.5%), mean CVMSE decreases from 46.92% to 41.79%, and mean R^2 increases from 0.926 to 0.948, while the worst-case error (Max RMSE) drops from 0.893 to 0.349 (−60.9%). For the Pathway layer, mean RMSE decreases from 0.203 to 0.153 (−24.6%), mean CVMSE decreases from 54.60% to 43.80%, and mean R^2 increases from 0.889 to 0.943, with worst-case RMSE reduced from 1.144 to 0.432 (−62.2%). Importantly, the added value is primarily reflected in tail-risk reduction (“bad-week” stability). In addition, the pipeline achieves these gains with a substantially more compact input set (Driver: 30 → 10; Pathway: 20 → 10), supporting that rank fusion yields a robust core feature subset rather than adding unnecessary complexity.

Table 5. Performance comparison: RF-only baseline vs. RRF+XGB (Top-10).

Layer	Model	K (#Features)	RMSE			CVMSE (Mean, %)	R^2 (Mean)
			Mean	Median	Max		
DRIVER	RF-only baseline	30	0.156	0.106	0.893	46.92	0.926
DRIVER	RRF+XGB (Top-10)	10	0.135	0.107	0.349	41.79	0.948
PATHWAY	RF-only baseline	20	0.203	0.103	1.144	54.60	0.889
PATHWAY	RRF+XGB (Top-10)	10	0.153	0.109	0.432	43.80	0.943

Note: RF-only uses the full feature set in each layer (DRIVER: 30; PATHWAY: 20). RRF+XGB uses Top-K features selected by RRF (K = 10). #: number of features.

The mean values across the rolling test sets (Driver layer) were RMSE = 0.1330, CVMSE = 40.7639%, and R^2 = 0.9500, indicating that the model could explain approximately 95% of the EUI variance without using future information, demonstrating strong overall fitting and extrapolation capabilities with relatively low absolute errors. The Pathway layer’s mean values were RMSE = 0.1477, CVMSE = 42.2171%, and R^2 = 0.9445, with

performance close to that of the Driver layer but with a slightly higher RMSE, suggesting that when the EUI is characterized by the energy channel variables, it may be more situation-dependent in some periods, thus leading to more error fluctuations. Overall, both layers maintained $R^2 \approx 0.95$ in the time extrapolation test, indicating that the main energy consumption change patterns could be stably captured, and the prediction had strong generalization ability. The RMSE was at a relatively low level, reflecting that the absolute error was controllable. It should be emphasized that the high mean value of CVRMSE does not imply prediction failure. CVRMSE is essentially $RMSE/\bar{y} \times 100\%$, and when the average load $\bar{y}$ of the test week is low, the percentage error will be passively magnified. Therefore, this paper does not solely use CVRMSE to determine the model's superiority or inferiority but uses it as a supplementary indicator of relative error and reports and interprets it in conjunction with RMSE and R^2 to obtain a comprehensive judgment of the model's accuracy and stability, thereby avoiding biased conclusions caused by load level changes on a single indicator. In the rolling extrapolation evaluation of the extended window, the average error and goodness of fit of RF in the Driver layer were $RMSE = 0.1348$, $CVRMSE = 41.78\%$, and $R^2 = 0.9469$. In the Pathway layer, they were $RMSE = 0.1552$, $CVRMSE = 45.02\%$, and $R^2 = 0.9333$. Both layers had R^2 exceeding 0.93, indicating that the model had a stable explanatory power for EUI fluctuations; at the same time, RMSE was at the 0.13–0.16 level, suggesting that the overall absolute prediction deviation was controllable. In contrast, RMSE and CVRMSE of the Pathway layer were slightly higher, indicating that under the influence of the Envelope Conduction, Occupancy and Lighting, and Infiltration channel variables, the impact of operation conditions and seasonal changes was more significant in some periods. Since CVRMSE normalizes $RMSE/\bar{y} \times 100\%$ by the mean value $\bar{y}$ of the test period, its cross-period comparison also needs to be interpreted in conjunction with RMSE and R^2 .

Based on the RRF ranking, the K sets for comparison were selected as follows: Driver layer: $K \in \{5, 10, 30\}$; Pathway layer: $K \in \{5, 10, 20\}$. It is worth emphasizing that this study observes clear method dependence in feature-importance ranking, i.e., the same feature can receive different importance under different importance methods.

In addition, this paper not only presents the rankings after fusion but also examines the stability and generalization performance of different Top- K truncations in the rolling evaluation. To avoid choosing K based only on visual inspection of RMSE/CVRMSE curves, the Top- K truncation is determined using a reproducible quantitative rule (Method 5) that combines CPSS stability selection, a stability-fraction gate, and a block-bootstrap 1SE performance gate. A stability-fraction gate is applied requiring $\text{Frac}(K) \geq \pi_0$. In this setting, the goal is not to minimize mean (CVRMSE%) alone, but to prefer a smaller, more interpretable feature set when its performance is not meaningfully worse than the best stability-eligible alternative under a bootstrap-1SE criterion; accordingly, “Final” is defined as the smallest K that passes both the stability gate and the bootstrap-1SE performance gate. The resulting K -selection outcome is summarized in Table 6.

Table 6. K-selection summary under Method 5 (CPSS stability-fraction + bootstrap-1SE).

Dataset	K	n_Windows	Mean CVRMSE (%)	Boot SE (%)	Stable Count	Stable Frac	Min π	Avg Rank	Win Rate	Final
driver	5	44	57.19	5.54	5	1.00	0.999	2.05	0.30	
driver	10	44	57.82	5.76	9	0.90	0.655	2.02	0.25	
driver	30	44	58.37	6.41	30	1.00	1.000	1.93	0.45	✓
pathway	5	44	98.96	9.55	5	1.00	0.993	2.95	0.00	
pathway	10	44	69.85	7.94	10	1.00	1.000	1.77	0.23	
pathway	20	44	65.54	8.80	20	1.00	1.000	1.27	0.77	✓

Notes: For the driver, best mean in stability set = 57.82% and bootstrap-1SE threshold = 63.58%. For the pathway, best mean in stability set = 65.54% and bootstrap-1SE threshold = 74.34%. Stable count/frac are computed on fullTopK using π from CPSS. ✓ indicates the final selected K under Method 5.

Table 6 summarizes the-selection outcome under Method 5 (CPSS stability selection+stability-fraction gate+block-bootstrap 1SE performance gate). The goal is not to minimize mean (CVRMSE%) alone, but to prefer a smaller, more interpretable feature set when its performance is not meaningfully worse than the best stability-eligible alternative under a bootstrap-1SE criterion.

For the driver dataset (with the pre-specified constraint $K \geq 10$), $K \geq 10$ achieves mean (CVRMSE%) = 57.82% and the bootstrap-1SE threshold is 63.58%; both $K = 10$ and $K = 30$ pass the performance gate, so the final rule selects the smallest eligible size, i.e., $K = 10$. For the pathway dataset, the lowest mean (CVRMSE%) occurs at $K = 20$ (65.54%), but $K = 20$ attains 69.85%, which remains within the 1SE threshold of 74.34%. Stability evidence is also comparably strong (stable fraction = 1.00 and min $\hat{\pi} = 1.000$), leading to the same final recommendation of $K = 10$.

Overall, the Top10 choice ($K = 10$) is not based on visual inspection of error curves. Instead, it follows a reproducible quantitative rule: the stability gate enforces consistency of the selected set across complementary resampling, while the bootstrap-1SE gate prevents adopting a larger, less interpretable set for only marginal gains in mean performance, thereby providing a statistically grounded justification for $K = 10$.

From the curve shapes in Figure 6, it can be seen that for the Driver layer, the rolling results show that the RMSE curve for Top-5 has higher peaks, indicating greater vulnerability to seasonal changes when there are fewer features; the Top-10 curve is overall smoother, demonstrating the best comprehensive performance; the Top-30 curve does not have a stable advantage overall. The CVRMSE curve also shows the same trend, with the Top-5 curve having greater fluctuations and more peaks; the Top-10 curve is the lowest and most stable overall, with the best robustness; the Top-30 curve is not systematically more stable than Top-10, and in some folds, the fluctuations are even more intense. Similarly, for R^2 , the Top-5 curve has deeper dips in some folds; the Top-10 curve is basically the most stable with the least fluctuations; the Top-30 curve does not outperform Top-10 and shows a diminishing return. Therefore, in the Driver layer, Top-10 strikes the best balance between accuracy and stability, Top-5 has larger fluctuations, and Top-30 does not show significant benefits and has increased fluctuation traces.

The rolling results of the Pathway layer from Figure 7 show that the RMSE curve of Top-5 has extremely prominent peaks, and the overall performance of the curve is not excellent; the Top-10 curve is the lowest and smoothest; Top-20 is not better than Top-10, and often has greater fluctuations. In the CVRMSE curve, the sharp peak of Top-5 may be related to the proportion amplification caused by the small mean energy consumption of this window, or it may be due to extreme weather or data issues. Top-10 is still the lowest and most stable, and performs extremely well in normal weeks; the characteristic of the Top-20 curve is that it returns to a low position and converges after the peak. In the R^2 curve, Top-5 has a cliff-like drop point, indicating that the data of that week is unstable; the Top-10 curve is still the most stable and optimal; after the drop point, the Top-20 curve returns to a relatively high level overall and tends to converge, indicating that the predictability of the Pathway feature is relatively high in normal weeks. In terms of RMSE, CVRMSE and R^2 indicators of the Driver layer and Pathway layer, Top-10 shows the best accuracy and stability. Compared with Top-5, RMSE and CVRMSE are lower and have less fluctuation, but a larger K shows a diminishing return and brings higher fluctuations or instability. The diminishing return of Top-30/20 has led to higher variations in several rounds. The results indicate that $K = 10$ is the best cut-off point in terms of comprehensive accuracy and robustness.

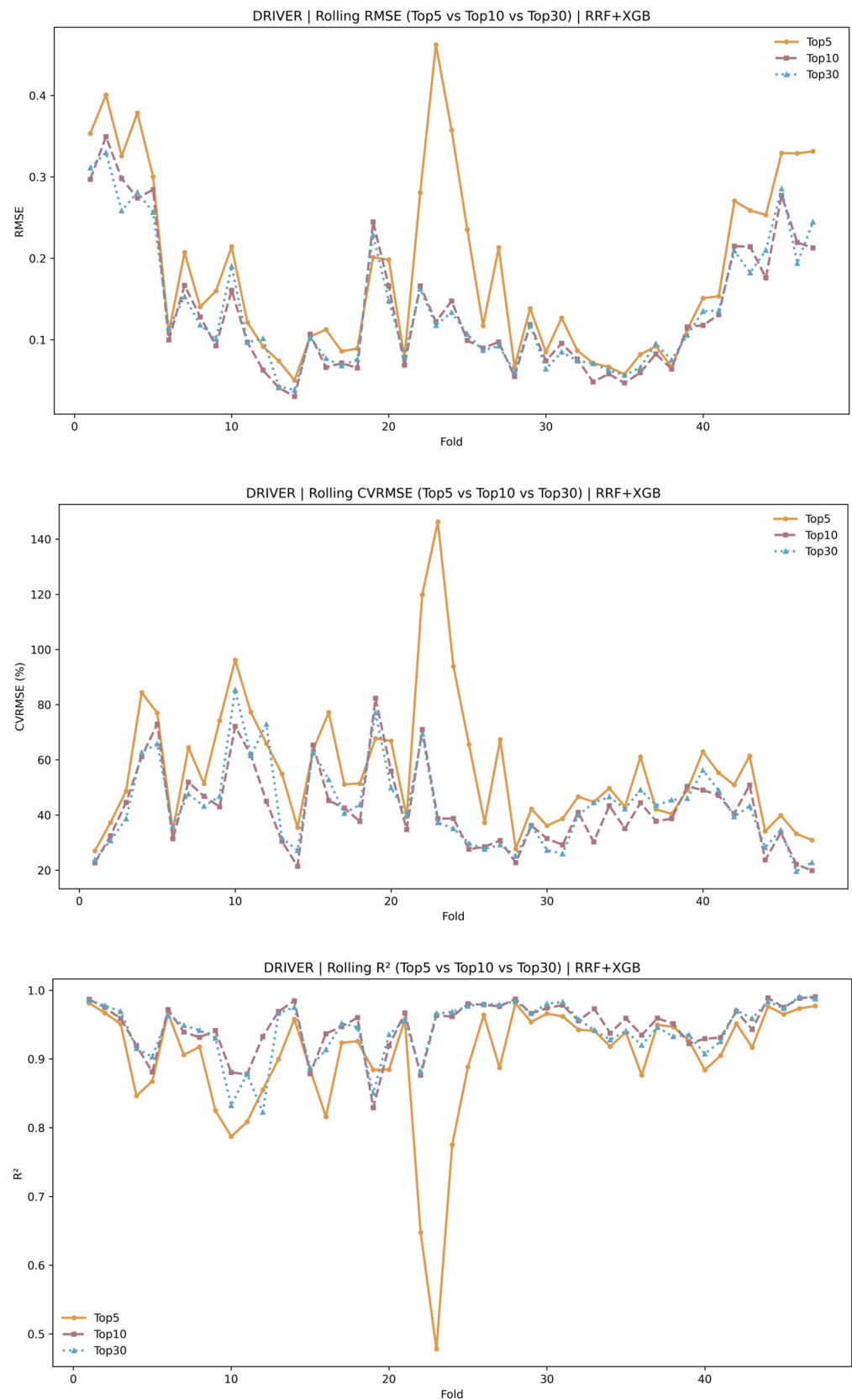


Figure 6. DRIVER rolling metrics across folds (Top-5/10/20 features).

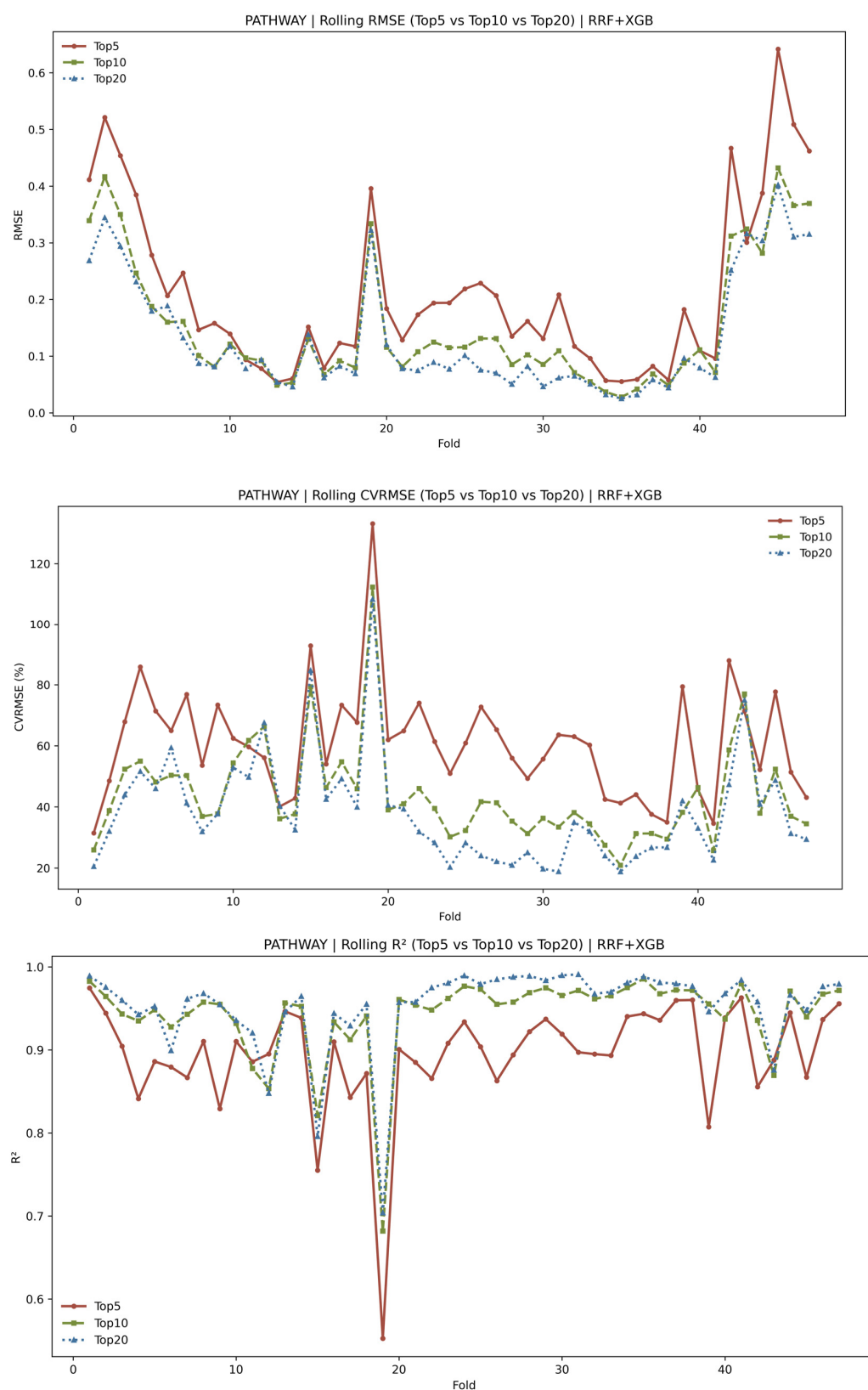


Figure 7. PATHWAY rolling metrics across folds (Top-5/10/20 features).

From Figures 8 and 9, it can be seen that each cell provides the feature importance ranking of the same feature under the three methods of Spearman, XGBoost-RFE, and Random Forest. The smaller the rank value, the more important the feature, and the darker the color, the higher the ranking. In Figure 8, the ranking differences in the same feature

in the Driver layer by different methods are obvious, indicating that this layer is more sensitive to the choice of methods. The differences may come from marginal correlation and model contribution, and nonlinearity and interaction are only one of the reasons for this difference. The Spearman and model-based XGBoost-RFE and RF have different focuses when capturing mechanisms, so the rankings of Spearman and XGBoost-RFE differ relatively more on several features. In Figure 9, the rankings in the Pathway layer are more consistent among different methods, indicating that the proximal energy channel-like features in this layer have a more direct and stable relationship with the target variable, and thus different methods are more likely to converge to a similar set of key features. Combining the two heatmap figures, it can be seen that the consistency of the top-ranked features across methods is more robust, while the consistency of the lower-ranked features decreases significantly and is more dependent on the method and data partitioning. This indicates that the ranking of a single method may have method dependence. To avoid this uncertainty, RRF is used to fuse the rankings of multiple methods and obtain a more robust ranking. At the same time, in the rolling evaluation, it is verified that the Top-10 features have the best balance between accuracy and stability.

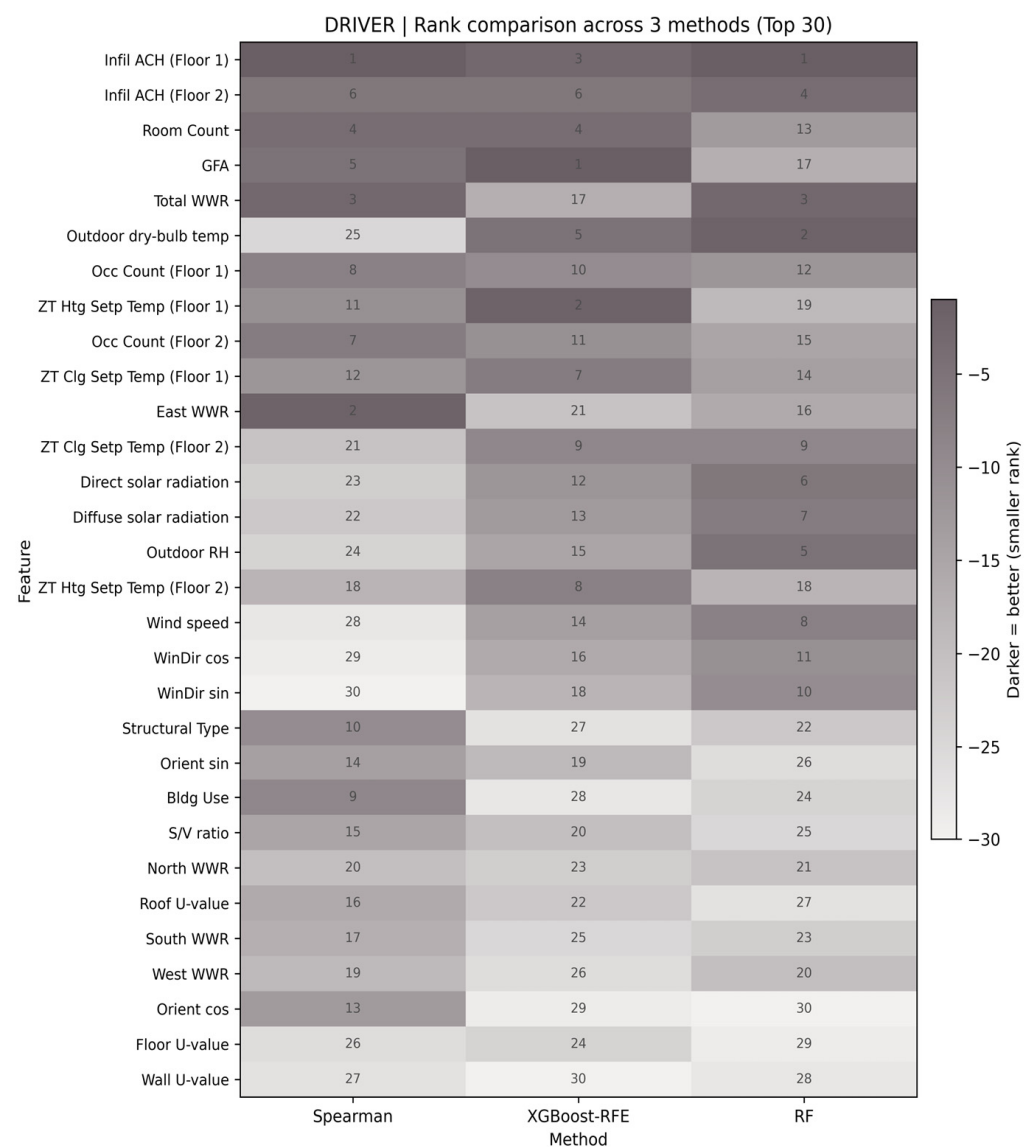


Figure 8. DRIVER feature ranking comparison across three methods (Top-30).

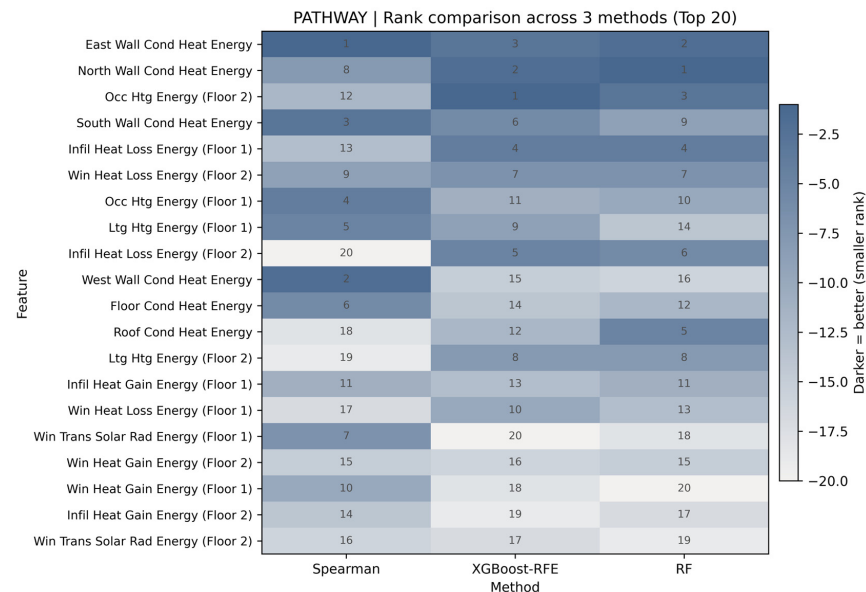


Figure 9. Pathway feature ranking comparison across three methods (Top-30).

4.3. SHAP-Based Interpretation Results and Discussion

This study employed three different approaches to feature selection: filter (Spearman), wrapper (XGBoost-RFE), and embedded (Random Forest). These methods can quantify the influence of input features on EUI, but they lack strong interpretability. Therefore, SHAP is needed to explain the impact of input features in model interpretation. Figure 10 provides a SHAP+XGBoost model analysis chart in descending order of feature importance. SHAP values are used to quantify and assess the contribution of each factor to the model's predictions. The x-axis shows the SHAP value for each factor: positive values increase the predicted EUI, whereas negative values decrease it. The y-axis lists the features ranked by importance. Each dot corresponds to an individual sample, and the color (red/blue) typically indicates the original feature value (red: high value; blue: low value). For ease of reading, only the Top-5 features are shown, which are the leading subset of the selected Top-K feature set.

From Figure 10a–f, the Driver layer SHAP summary and dependence plots (Top-5) indicate that the model output is most strongly associated with infiltration-related operational predictors, with Infil ACH (Floor 1) showing the largest mean absolute SHAP value (Figure 11a). Its point cloud extends predominantly into the positive SHAP range, and higher ACH values are consistently linked to higher predicted EUI, suggesting a pronounced nonlinear sensitive range at medium-to-high infiltration levels (Figure 10b). GFA and Infil ACH (Floor 2) exhibit secondary contributions with clear dispersion and overlap, implying value-dependent effects and potential interaction with other operational conditions (Figure 10c–d). ZT Htg Setp Temp (Floor 1) shows a comparatively smaller but directionally meaningful contribution, reflecting the sensitivity of predictions to thermostat control settings under different operating regimes (Figure 10e). In contrast, Room Count has a narrow SHAP amplitude around zero and appears to function mainly as a weak proxy for building form/functional complexity, with much of its explanatory content likely absorbed by more directly operational and size-related variables (Figure 10f). Overall, the Driver layer SHAP patterns emphasize operational-control sensitivity rather than macro-scale structural heterogeneity. From Figure 11a–f, it can be seen that the SHAP summary plot (Top-5) of the Pathway layer indicates that the model's predictions are mainly associated with envelope conduction and infiltration-related heat-balance terms. Specifically, North/South/East Wall Cond Heat Energy exhibit comparatively wide SHAP spreads across samples (Figure 11a),

suggesting a stable and non-negligible contribution of the envelope conduction pathway to the predicted EUI within the calibrated EnergyPlus context. The dependence plots further reveal nonlinear and threshold-like behaviors. For example, the SHAP response to wall conduction terms is not strictly linear (Figure 11b,d,e), implying value-dependent sensitivities and potential coupling with other boundary-condition variables. Occ Htg Energy (Floor 2) shows a clear regime-dependent pattern (Figure 11c): its contribution can shift in magnitude and even direction, consistent with a “mode-switching” effect of internal gains under seasonal operation. Infil Heat Loss Energy (Floor 1) demonstrates a distinct nonlinear sensitive range (Figure 11f), where higher infiltration losses correspond to rapidly increasing positive SHAP values, indicating that intensified infiltration-related losses are strongly associated with higher predicted EUI. Overall, these SHAP patterns provide physics-informed interpretation of dominant loss channels rather than causal proof.

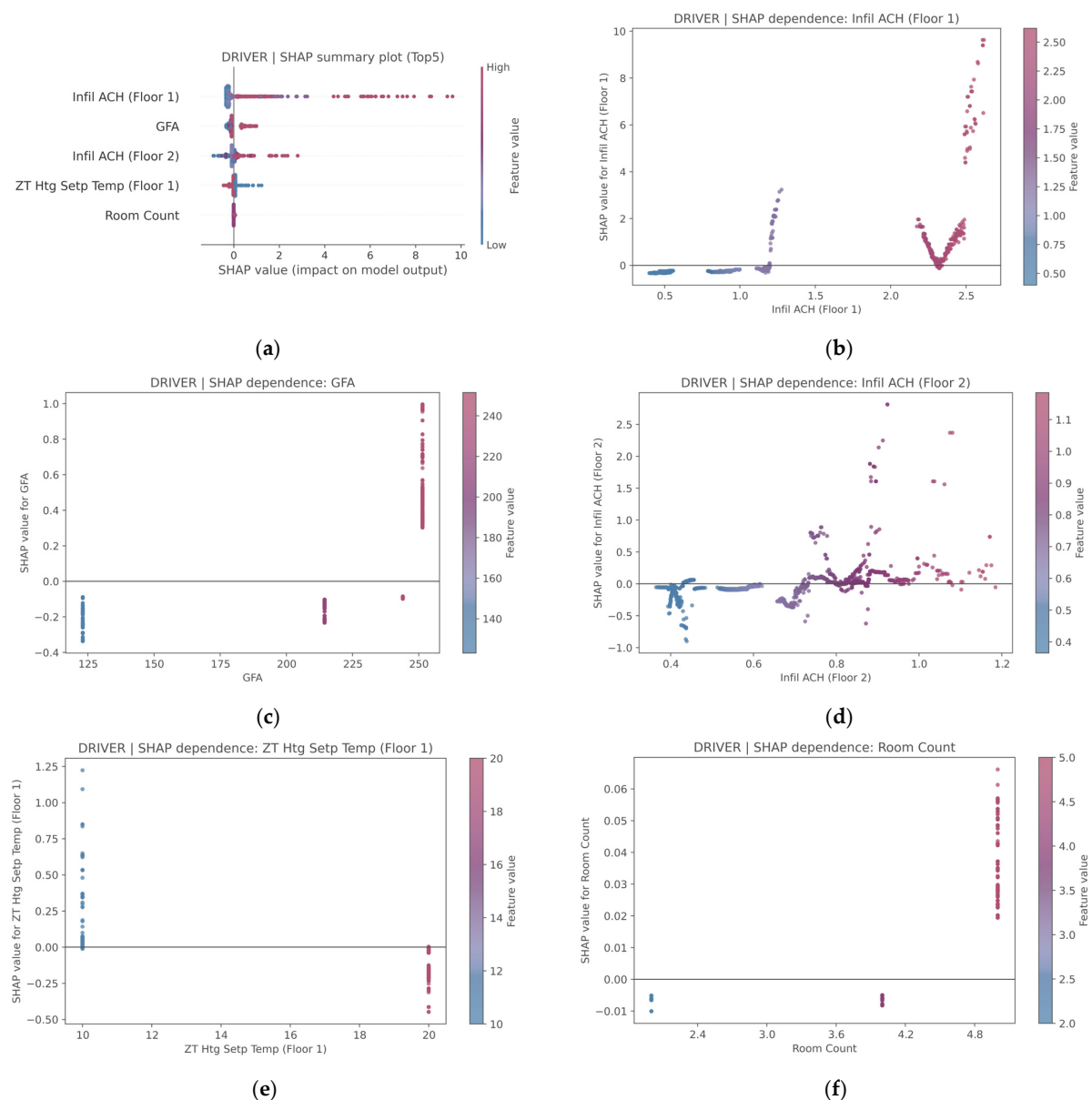


Figure 10. DRIVER SHAP summary and dependence plots (Top-5): (a) SHAP summary plot of the top five features; (b) SHAP dependence plot for Infil ACH (Floor 1); (c) SHAP dependence plot for GFA; (d) SHAP dependence plot for Infil ACH (Floor 2); (e) SHAP dependence plot for ZT Htg Setp Temp (Floor 1); (f) SHAP dependence plot for Room Count.

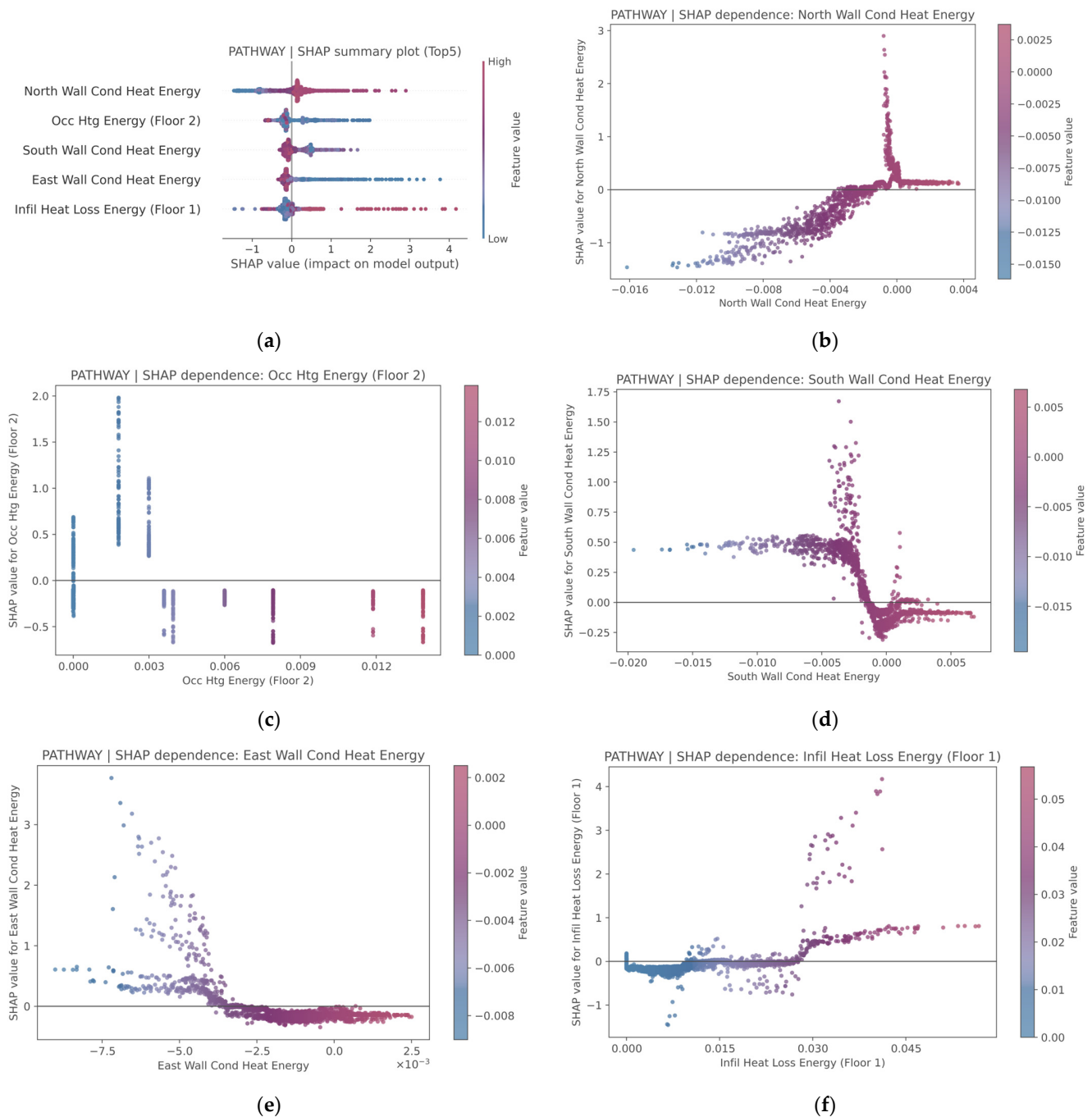


Figure 11. PATHWAY SHAP summary and dependence plots (Top-5): (a) SHAP summary plot of the top five features; (b) SHAP dependence plot for North Wall Cond Heat Energy; (c) SHAP dependence plot for Occ Htg Energy (Floor 2); (d) SHAP dependence plot for South Wall Cond Heat Energy; (e) SHAP dependence plot for East Wall Cond Heat Energy; (f) SHAP dependence plot for Infil Heat Loss Energy (Floor 1).

5. Discussion

5.1. Dominant Drivers and Pathway Structure Under Controlled Comparability

After the similarity screening, the findings of this study suggest that EUI differences among historic buildings are no longer mainly dominated by the heterogeneity of macro-level structures, but are concentrated in operational drivers (such as Infiltration, Operational Setpoints, Occupancy) and their corresponding load formation channels (Envelope Conduction, Infiltration/Windows, Internal heat gain). This finding supports the working

hypothesis of this paper: when the research objects have strong comparability in specific aspects such as Building form and Energy carrier, factors such as occupant behavior and operational control can have an explanatory power for energy consumption differences comparable to that of building physical attributes, which makes the explanation closer to intervenable factors. At the same time, a relatively clear mechanism loop is formed between the Driver layer and the Pathway layer results: the top-ranked intervenable features in the Driver layer, such as Infil ACH and ZT Htg/Clg Setp Temp, usually correspond to the main loss channels such as Infil Heat Loss Energy, Wall Cond Heat Energy and Win Heat Loss Energy in the Pathway layer. The key features not only remain highly ranked in a statistical sense, but also enable their effects to be traced to specific heat-loss pathways through machine learning methods, thereby providing a mechanistic explanation.

5.2. Ranking Robustness: Method Dependence, RRF, and Top-K Selection

It is also worth emphasizing that this study observed certain differences in the importance of the same feature under different importance methods, suggesting that the importance ranking of a single method may be method-dependent. A comparison of the three methods (Figures 9 and 10) reveals that, on the one hand, features related to infiltration (such as Infil ACH) and several dominant heat channels (Wall Cond Heat Energy, Infil Heat Loss Energy, Win Heat Loss Energy) are consistently ranked high across different methods, indicating that these features can be regarded as more robust and reliable key features. On the other hand, features such as Solar Radiation, Wind direction, and U-Value show significant fluctuations under different methods, suggesting that their effects may be more dependent on nonlinear, interactive, or coupled relationships with other features. Therefore, it is not advisable to draw qualitative conclusions based solely on the ranking of a single method. Based on these results that are both consistent and divergent, this paper further employs RRF to fuse the rankings of multiple methods to reduce the bias of a single method and enhance the robustness of identifying key features.

In addition, this paper not only presents the rankings after fusion but also examines the stability and generalization performance of different Top-K truncations in the rolling evaluation: the results show that Top-10 achieves a better trade-off between accuracy and stability (with lower error and less fluctuation compared to Top-5), while larger K values generally show diminishing returns and may introduce higher volatility. This conclusion has direct methodological implications for subsequent research: in the context of similar building comparisons and cross-period tests, rather than continuously expanding the feature set for full coverage, it is better to prioritize the identification of key features that are stable across multiple methods and periods, thereby obtaining reproducible features and facilitating the formation of interpretable intervention priority results.

Furthermore, in the Pathway layer, the consistency of Spearman is relatively poor, indicating that Spearman may have limitations in identifying the importance of heat loss pathway features. A more reasonable explanation is that the output of the Pathway layer often contains significant seasonal switching, feature changes, and nonlinear coupling, while Spearman mainly characterizes monotonic relationships. Therefore, it is more likely to underestimate the contributions of those non-monotonic but crucial pathways, leading to insufficient stability in the ranking. Based on this observation, this paper suggests that subsequent studies should avoid relying solely on correlation coefficient-based methods in Pathway layer importance analysis. Instead, they should combine model-driven methods capable of characterizing nonlinearity and interaction (such as model-based feature importance, RFE, and SHAP) as well as multi-method fusion and cross-period rolling validation to enhance the reliability and reproducibility of channel identification.

5.3. Intervention Implications and Prioritization

5.3.1. Six-Category Interpretation and Motivation

In order to transform the quantitative results of this paper into interpretable mechanism paths and actionable intervention inspirations, this paper further starts from the six types of building energy consumption influencing factors commonly adopted in the academic circle, and classifies and discusses the key factors identified in the driver layer and pathway layer. This framework not only helps to clarify the mechanism of action of different factors on EUI, but also facilitates the distinction between intervention and renovation risks under the protection constraints of historical buildings, thereby forming a more executable priority for energy-saving strategies. Accordingly, the following discussion is organized by the six categories, linking the Driver layer and Pathway layer findings to mechanism interpretation and conservation-compatible intervention priorities.

1. Climate

Climate is not a direct intervention target. It is an external boundary condition for the formation of building energy consumption. These variables themselves cannot be changed, but they play a key role in explaining the driving factors of EUI, mainly for energy consumption normalization and robustness tests (such as rolling validation), to avoid misjudging the differences between cold and hot years as differences in building or operation. From the perspective of energy conservation, the climate boundary will amplify the energy consumption consequences of infiltration, window and door losses, and control strategies through mechanisms such as solar heat gain, wind speed/direction and wind-driven pressure difference, and moisture load. Therefore, at the strategy level, climate-responsive operation should be emphasized to achieve energy conservation goals. Consistent with the Driver layer rank comparison across methods, outdoor dry-bulb temperature remains a top-ranked climate variable (Spearman: 2; XGBoost-RFE: 5; RF: 2), indicating that climate is a non-negligible boundary condition for EUI interpretation even after controlling for building and operational differences.

2. Building envelope

Building envelope is a factor with a low degree of intervention. The renovation of the envelope of historical buildings is restricted by conservation constraints. Energy conservation for this factor should not involve large-scale alterations. Priority should be given to small-scale, reversible measures targeting the dominant heat loss channels identified in the pathway layer, such as wall conduction heat energy, infiltration heat loss energy, and window heat loss energy. For example, air tightness repair of gaps around door and window openings, entrance vestibules or air curtains, reversible window and door enhancements (secondary glazing), and local low-risk interior insulation. At the same time, it must be advanced in parallel with the control of wet damage risks, and the coupling amplification effect of heat load and moisture load should be reduced by avoiding uncontrolled infiltration. Therefore, the energy conservation strategies for this category of factors can be carried out following the principles of channel orientation, targeted intervention, and priority for reversible renovations. Consistent with the Pathway layer rank comparison, wall conduction pathways remain dominant (e.g., East Wall Cond Heat Energy: 1/3/2; North Wall Cond Heat Energy: 8/2/1; South Wall Cond Heat Energy: 3/6/9 across Spearman/XGBoost-RFE/RF), suggesting that envelope conduction constitutes a primary load formation mechanism driving EUI variation.

3. Building services and energy systems (Services and energy systems)

Building services and energy systems (Services and energy systems) fall into the category of factors with a medium level of intervention, typically being more reversible and

having a lesser impact on the historical building itself compared to envelope renovations. They are thus suitable as one of the preferred technical approaches under the constraints of historical building conservation. Based on the variable system and result evidence presented in this article, energy-saving suggestions mainly manifest in operational control and terminal treatment. This includes optimizing operation periods and zoning strategies, integrating control logics, and upgrading the efficiency and control of lighting systems, such as using LEDs, zoning, sensors, and time control. Lighting, in particular, not only has the advantages of strong reversibility and minimal implementation disruption but also affects the heating and cooling loads through internal heat gain. Therefore, it is recommended to evaluate it in conjunction with other strategies to avoid frequent switching of operation mechanisms due to single-point optimization. Before drawing the conclusion of this article, it should be noted that, although HVAC is not included as an explanatory variable in this article, and given that plug-load and equipment end uses are not introduced as predictive features (see Table 1), no quantitative conclusions are provided for specific equipment renovations. This part will be left for further verification in the future when more detailed data is available. Consistent with the Pathway layer rank comparison, lighting-related pathways remain within the top set (e.g., Ltg Htg Energy (Floor 1): 5/9/14; Ltg Htg Energy (Floor 2): 19/8/8), indicating that lighting-related internal heat gains (and associated control/operation patterns) are strongly associated with EUI differences in the sample.

4. Building operation and maintenance (O&M)

Building operation and maintenance (O&M) belong to the category of factors with a high level of intervention, strong short-term operability, and the lowest cost. These factors often influence EUI through operation periods, start-up strategies, ventilation modes, equipment interlocking logics, and management systems, and can significantly explain why there are still significant energy consumption differences among similar buildings when the envelope cannot be significantly enlarged. It is recommended to prioritize low-cost operational optimizations, including operation periods and zoning strategies, ventilation operation, pressure difference control, system commissioning, and setpoint strategies that match occupancy, to achieve rapid and verifiable energy savings. Consistent with the Driver layer rank comparison, Bldg Use appears within the Top-30 across all three methods (Spearman: 9; XGBoost-RFE: 28; RF: 24), supporting that use-type-driven operation patterns and schedules are material, intervenable contributors to EUI variation under conservation constraints.

5. Occupant activities and behavior (Occupant behavior)

Occupant activities and behavior (Occupant behavior) fall into a category of medium to high intervention levels. Due to the significant time-varying nature of human behavior, the number of occupants, their stay time, and activity intensity directly determine the internal heat gain and the fluctuation of instantaneous loads. Human behaviors such as door opening frequency, window opening and shading, lighting and equipment usage habits can be through behavioral interventions and control strategies. However, it needs to be discussed in conjunction with building use. Different business types correspond to different business hours, customer flow density, and indoor environmental demands, which determine the reasonable boundaries of behavioral patterns and control strategies. It is suitable to prioritize people-centered strategies, promoting the joint optimization of occupancy, ventilation, and set points. For instance, dynamically adjust the ventilation volume according to the occupancy level during operation periods, and based on the robustness results of rolling evaluations, prioritize stable coordinated strategies to reduce the risk of inefficient operation and achieve energy conservation. Consistent with the Pathway layer rank comparison, occupant-related internal-gain pathways rank among

the very top (Occ Htg Energy (Floor 2): 1/2/3; Occ Htg Energy (Floor 1): 4/11/10), underscoring that occupant-driven internal heat gains are tightly linked to EUI variation and should be addressed via occupant-centric operation.

6. Indoor environmental quality provided (IEQ)

Indoor environmental quality provided (IEQ) belongs to a category with high intervention but under constraints. It is closer to a constraint variable rather than a result variable: set points, thermal comfort and air quality requirements directly limit the available energy-saving space. Therefore, energy-saving strategies in this category should not be simple measures such as lowering heating set points or raising cooling set points. Instead, multi-objective trade-offs should be made within the boundaries of comfort and air quality. For example: within the acceptable range, expand the non-conditioning (deadband) periods, adopt seasonal set points and time-of-day strategies, implement occupancy-matched ventilation control, and ensure that energy savings do not come at the expense of environmental quality by monitoring temperature, humidity and indoor air quality (IAQ) indicators. The minimum energy consumption achieved through multi-objective trade-offs under constraints is the approach to energy conservation.

5.3.2. SHAP-Based Mechanistic Insights and Prioritization

From the perspective of mechanistic interpretation, the SHAP dependence results complement the ranking with key information such as directionality, thresholds, and interactions. Infiltration-related features exhibit a distinct nonlinear sensitive range: at higher infiltration levels, the marginal contribution of the feature to the prediction increases more rapidly, indicating that unorganized air infiltration disproportionately amplifies heating/cooling and dehumidification loads. This is consistent with the common air tightness defects, door and window openings, and pressure-driven infiltration mechanisms in historic buildings. Occ Htg Energy shows a threshold effect and nonlinear behavior, and its net contribution to the prediction may change sign (from positive to negative or vice versa). This indicates a “mode-switching effect” of internal heat gain: in the heating season, it may manifest as supplementary heat benefits, while in the cooling season, it may transform into an accumulation of sensible heat and dehumidification loads, and in high occupancy situations, it may interact with exhaust ventilation to amplify the system load. ZT Htg Setp Temp shows a clear clustering feature, further suggesting that its effect is highly dependent on outdoor conditions and operating regimes, making it a typical feature with strong interactions. Therefore, the results of this study are more inclined to support priority judgments based on mechanistic explanations rather than simply equating the ranking results with linear causal strengths. A more reasonable approach is to prioritize infiltration (because it is consistently prominent in both the Driver and Pathway layers and has a high-sensitivity range), followed by focusing on the Envelope Conduction dominant loss channel, and treating occupant behavior and operational control as objects for coordinated optimization to avoid the switching of energy consumption mechanisms caused by internal heat gain under different conditions.

Based on the above-mentioned nonlinear, threshold and operational mode switching characteristics, the results of this paper can be further transformed into intervention priority implications for historic buildings. Firstly, infiltration-related features are consistently prominent in both the Driver layer and Pathway layer, indicating that airtightness and ventilation organization are the most robust and interpretable priority features. Under the relevant protection constraints of historic buildings, priority should be given to reversible operational control changes and detailed restoration measures, and combined with pressure difference and fresh air organization in operations to avoid unorganized infiltration taking on the ventilation function and amplifying the thermal and moisture loads. For example,

the competent authorities can reduce the rigid demand for high outdoor-air rates and precise humidity control through functional positioning and usage intensity management (such as choosing commercial, residential, and other business types that are more suitable for historic buildings), thereby reducing energy consumption and protection conflicts from the source. This can reduce building energy consumption at the management level, thereby alleviating financial pressure and improving the economic sustainability of historic building protection. Secondly, the Pathway layer shows that Wall Cond Heat Energy and Infil Heat Loss Energy constitute the dominant loss channels, suggesting that the insulation strategy should focus more on mitigating the key heat loss paths (such as treating walls and orientations with more significant heat loss), rather than pursuing uniform upgrades for all components. Thirdly, internal heat gains (Occupancy and Lighting) indicate that related strategies need to be coordinated with seasons and operation schedules: in situations dominated by heating and cooling, the net effect of internal heat gains on EUI may be opposite. Therefore, a more reasonable direction is to pursue coordinated control of the occupancy-ventilation-setpoint interaction, rather than assuming that reducing internal heat gains will necessarily reduce EUI in all periods. Overall, the two-layer framework of this paper establishes a mapping between the operational Driver layer features and the interpretable Pathway layer channels, enabling energy-saving suggestions to shift from empirical to priority judgments under mechanism constraints, thereby enhancing the reliability and transferability of intervention prioritization.

5.3.3. Implementation Guidance

Accordingly, managers and design/operation teams can apply the proposed process to: (1) prioritize low-risk, reversible, and readily adjustable operational-control actions without altering historic construction (e.g., schedules, setpoints, ventilation strategy and pressure-difference control, commissioning/re-commissioning), and use rolling evaluation to test the stability of strategies across periods; (2) when envelope or opening-related measures are truly necessary, implement targeted mitigation based on the dominant heat-loss pathways identified in the Pathway layer, while assessing moisture-damage risks in parallel and avoiding unorganized infiltration that may amplify coupled heat and moisture loads; (3) within the policy context of “promoting conservation through utilization”, treat building use (business-type configuration) as an actionable lever: differences in operating hours, occupancy density and variability, door-opening frequency, ventilation requirements, and thermal-comfort targets across business types can systematically reshape EUI, so leasing and use-type management, zoning, and the matching of control strategies to use intensity and comfort demands can form more verifiable and lower-risk energy-saving routes than high-intervention retrofits. In other words, this paper shifts decision-making from an experience-based assumption of “envelope retrofit first” toward a mechanism-oriented sequence of “diagnose dominant channels first, then apply targeted interventions”, thereby reducing decision uncertainty before capital investment and avoiding unnecessary engineering input and potential conservation risks.

6. Conclusions

This study is positioned as a methodological demonstration under the constraints of historic building preservation. Its core contribution does not lie in providing directly extrapolatable “general conclusions”, but rather in offering a replicable and verifiable framework that links operational drivers to explainable pathways to guide conservation-compatible intervention priorities. It is important to note that the fused ranking of key drivers/features and the dominant patterns in the Pathway layer reported in this study were derived from four in-use historic buildings in Shenyang under specific climate boundary

conditions and operational conditions, following similarity screening and assessed through cross-period rolling evaluation (Rolling Validation). Therefore, the identified rankings and Pathway layer patterns are context-dependent and may vary with climate years, building use (business types), operational strategies, occupant behavior, and existing renovation status. Based on this, this study does not advocate directly extrapolating the current driver ranking and Pathway layer patterns to other cities, other climate zones, other building-use categories/business types, or other building forms (e.g., non-rectangular, polygonal, or L-shaped plans), as well as row (terraced) buildings attached to adjacent new constructions, without additional validation.

It should be noted that the conclusions of this paper mainly come from correlational evidence under the model interpretation framework and cannot be directly equated with strict causal effects. Some proxy variables may bring measurement errors and comparability residuals. The Pathway layer channel variables have obvious seasonal switching and nonlinear coupling, leading to consistency differences among different importance methods. Although this paper adopts multi-method fusion and rolling validation to improve robustness, it is still difficult to completely eliminate dependence on the method and the model. Additionally, although the model has been calibrated and verified, the room temperature under the simulation conditions is the actual indoor temperature of the building and does not consider the requirements of thermal comfort. If the setpoint is adjusted to meet thermal comfort requirements, the energy consumption is expected to increase further.

Nevertheless, the results of this paper still provide clear practical usability. In past energy-saving practice for historic buildings, it has often been implicitly assumed that improvements mainly rely on envelope retrofits or physical interventions to the building fabric. However, under conservation constraints, such measures are not only costly and associated with substantial approval and construction disturbance, but may also introduce condensation and moisture-damage risks, making the trade-offs between “investment-benefit” and “energy saving-conservation” more sensitive. The “Driver-Pathway-EUI” mechanistic chain and the six-category factor framework proposed in this study, together with multi-method ranking fusion and rolling evaluation/rolling validation, can answer a more decision-relevant question before intervention: which operable driver factors and their corresponding load-formation channels mainly dominate the observed EUI differences.

Regarding excluded factors and future development, this study does not perform disaggregated modeling of end uses such as equipment/plug loads under the current data conditions and research objectives. This does not imply that they are unimportant; rather, without sub-metering and finer-grained business-type/equipment data, it is difficult to derive verifiable equipment-level conclusions at this stage. With access to end-use metering, refined business-type grouping, and more detailed operational records, such end-use variables can be incorporated to improve the explanatory precision for internal heat gains and end-use loads. In addition, the rolling evaluation in this paper has provided a basis for assessing cross-period robustness; subsequent work can further refine scenario and strategy tests along temperature and time dimensions, while maintaining the principle of preventing future information leakage, to complement the discussion of strategy adaptability across seasons and operating periods. At the sample level, the next step will build on the same screened pool and conduct multi-case replication and external validation on additional standalone buildings within the same selection criteria, to test the transferability of the framework within comparable objects; thereafter, the study can be extended to other cities and climate zones for cross-regional validation, so as to systematically characterize the applicability boundary and generalization conditions of the proposed methodology. Intervention priorities suggested in this study should be confirmed through controlled

experimental or quasi-experimental designs, such as A/B testing of operational parameters (e.g., thermostat setpoints or ventilation control strategies) or pre- and post-modernization monitoring with weather- and occupancy-normalized energy data, ideally with a matched control building or zone.

Author Contributions: Data curation, C.L. and Q.Z.; Formal analysis, C.L.; Methodology, C.L. and F.L. and Q.Z.; Project administration, F.L.; Supervision, F.L. and Q.Z.; Writing—original draft, C.L.; Writing—review and editing, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China (No. 52578005).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Annibaldi, V.; Cucchiella, F.; De Berardinis, P.; Gastaldi, M.; Rotilio, M. An Integrated Sustainable and Profitable Approach of Energy Efficiency in Heritage Buildings. *J. Clean. Prod.* **2020**, *251*, 119516. [\[CrossRef\]](#)
2. Zhao, H.; Magoulès, F. A Review on the Prediction of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3586–3592. [\[CrossRef\]](#)
3. Zhang, T.; Tan, Y.; Yang, H.; Zhang, X. The Application of Air Layers in Building Envelopes: A Review. *Appl. Energy* **2016**, *165*, 707–734. [\[CrossRef\]](#)
4. Masoso, O.T.; Grobler, L.J. The Dark Side of Occupants' Behaviour on Building Energy Use. *Energy Build.* **2010**, *42*, 173–177. [\[CrossRef\]](#)
5. Andargie, M.S.; Touchie, M.; O'Brien, W. A Review of Factors Affecting Occupant Comfort in Multi-Unit Residential Buildings. *Build. Environ.* **2019**, *160*, 106182. [\[CrossRef\]](#)
6. Janda, K. Buildings Don't Use Energy: People Do. *Archit. Sci. Rev.* **2011**, *54*, 15–22. [\[CrossRef\]](#)
7. Zhou, N.; Khanna, N.; Feng, W.; Ke, J.; Levine, M. Scenarios of Energy Efficiency and CO₂ Emissions Reduction Potential in the Buildings Sector in China to Year 2050. *Nat. Energy* **2018**, *3*, 978–984. [\[CrossRef\]](#)
8. You, K.; Ren, H.; Cai, W.G.; Huang, R.; Li, Y. Modeling Carbon Emission Trend in China's Building Sector to Year 2060. *Resour. Conserv. Recycl.* **2023**, *188*, 106679. [\[CrossRef\]](#)
9. Yu, Y.; You, K.; Cai, W.; Feng, W.; Li, R.; Liu, Q.; Chen, L.; Liu, Y. City-Level Building Operation and End-Use Carbon Emissions Dataset from China for 2015–2020. *Sci. Data* **2024**, *11*, 138–153. [\[CrossRef\]](#)
10. Ni, Z.; Eriksson, P.; Liu, Y.; Karlsson, M.; Gong, S. Improving Energy Efficiency While Preserving Historic Buildings with Digital Twins and Artificial Intelligence. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *863*, 12041. [\[CrossRef\]](#)
11. Cumo, F.; Nardecchia, F.; Agostinelli, S.; Rosa, F. Transforming a Historic Public Office Building in the Centre of Rome into nZEB: Limits and Potentials. *Energies* **2022**, *15*, 697. [\[CrossRef\]](#)
12. Wise, F.; Moncaster, A.; Jones, D.; Dewberry, E. Considering Embodied Energy and Carbon in Heritage Buildings—A Review. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *329*, 12002. [\[CrossRef\]](#)
13. Streicher, K.N.; Padey, P.; Parra, D.; Bürer, M.C.; Schneider, S.; Patel, M.K. Analysis of Space Heating Demand in the Swiss Residential Building Stock: Element-Based Bottom-up Model of Archetype Buildings. *Energy Build.* **2019**, *184*, 300–322. [\[CrossRef\]](#)
14. Xu, C.; Gong, A.; Liang, L.; Song, X.; Wang, Y. Vulnerability Assessment Method for Immoveable Cultural Relics Based on Artificial Neural Networks—An Example of a Heavy Rainfall Event in Henan Province. *Int. J. Disaster Risk Sci.* **2023**, *14*, 41–51. [\[CrossRef\]](#)
15. Wang, Y.; Zhen, Y.; Ning, Q. Four-Objective Optimization of Historic Building Renovation Considering Climate Change: A Trade-off in Energy, Economy, Environment, and Thermal Comfort. *J. Build. Eng.* **2025**, *110*, 113017. [\[CrossRef\]](#)
16. Mohamed, B.; Marzouk, M. Post-Adaptive Reuse Evaluation of Heritage Buildings Using Multi-Criteria Decision-Making Techniques. *J. Build. Eng.* **2025**, *99*, 111485. [\[CrossRef\]](#)
17. Zhang, Y.; Zhang, Q. A Model Approach for Post Evaluation of Adaptive Reuse of Architectural Heritage: A Case Study of Beijing Central Axis Deciding on the selection of features require Buildings. *Herit. Sci.* **2023**, *11*, 57–74. [\[CrossRef\]](#)
18. Gravagnuolo, A.; Angrisano, M.; Bosone, M.; Buglione, F.; De Toro, P.; Fusco Girard, L. Participatory Evaluation of Cultural Heritage Adaptive Reuse Interventions in the Circular Economy Perspective: A Case Study of Historic Buildings in Salerno (Italy). *J. Urban Manag.* **2024**, *13*, 107–139. [\[CrossRef\]](#)
19. Boccardi, G. Authenticity in the Heritage Context: A Reflection beyond the Nara Document. *Hist. Environ. Policy Pract.* **2019**, *10*, 4–18. [\[CrossRef\]](#)

20. Vicha, O. The Concept of the Right to Cultural Heritage within the Faro Convention. *Int. Comp. Law Rev.* **2014**, *14*, 25–40. [\[CrossRef\]](#)
21. Della Spina, L.; Carbonara, S.; Stefano, D.; Viglianisi, A. Circular Evaluation for Ranking Adaptive Reuse Strategies for Abandoned Industrial Heritage in Vulnerable Contexts. *Buildings* **2023**, *13*, 458. [\[CrossRef\]](#)
22. Dişli, G.; Ankaralıgil, B. Circular Economy in the Heritage Conservation Sector: An Analysis of Circularity Degree in Existing Buildings. *Sustain. Energy Technol. Assess.* **2023**, *56*, 103126. [\[CrossRef\]](#)
23. Foster, G. Circular Economy Strategies for Adaptive Reuse of Cultural Heritage Buildings to Reduce Environmental Impacts. *Resour. Conserv. Recycl.* **2020**, *152*, 104507. [\[CrossRef\]](#)
24. Marzouk, M. *Heritage Building Conservation: Sustainable and Digital Modelling*, 1st ed.; Routledge: London, UK, 2023.
25. Walker, R.; Pavia, S. Thermal and Moisture Monitoring of an Internally Insulated Historic Brick Wall. *Build. Environ.* **2018**, *133*, 178–186. [\[CrossRef\]](#)
26. Raffler, S.; Bichlmair, S.; Kilian, R. Mounting of Sensors on Surfaces in Historic Buildings. *Energy Build.* **2015**, *95*, 92–97. [\[CrossRef\]](#)
27. Liu, Y.; Akram Hassan, K.; Karlsson, M.; Pang, Z.; Gong, S. A Data-Centric Internet of Things Framework Based on Azure Cloud. *IEEE Access* **2019**, *7*, 53839–53858. [\[CrossRef\]](#)
28. Liu, Y.; Akram Hassan, K.; Karlsson, M.; Weister, O.; Gong, S. Active Plant Wall for Green Indoor Climate Based on Cloud and Internet of Things. *IEEE Access* **2018**, *6*, 33631–33644. [\[CrossRef\]](#)
29. Bartolucci, B.; Frasca, F.; Flores-Colen, I.; Bertolin, C.; Siani, A.M. Key Performance Indicators: Their Use in the Energy Efficiency Retrofit for Historic Buildings. *Procedia Struct. Integrity* **2024**, *55*, 110–118. [\[CrossRef\]](#)
30. Lo Faro, A.; Nocera, F.; Taranto, V. The Dynamic Thermal Energy Simulation of Historic Buildings in Mediterranean Climate: Knowledge and Simplification. *IOP Conf. Ser. Earth Environ. Sci* **2021**, *863*, 012011. [\[CrossRef\]](#)
31. Frasca, F.; Cornaro, C.; Siani, A.M. A Method Based on Environmental Monitoring and Building Dynamic Simulation to Assess Indoor Climate Control Strategies in the Preventive Conservation within Historical Buildings. *Sci. Technol. Built Environ.* **2019**, *25*, 1253–1268. [\[CrossRef\]](#)
32. Cassar, M.; Pender, R. The impact of climate change on cultural heritage: Evidence and response. In *ICOM Committee for Conservation: 14th Triennial Meeting the Hague, Preprints*; James & James: Kamakura, Japan, 2005; Volume 2, pp. 610–616.
33. Liu, X.; Tang, H.; Ding, Y.; Yan, D. Investigating the Performance of Machine Learning Models Combined with Different Feature Selection Methods to Estimate the Energy Consumption of Buildings. *Energy Build.* **2022**, *273*, 112408. [\[CrossRef\]](#)
34. Neubauer, A.; Brandt, S.; Kriegel, M. Relationship between Feature Importance and Building Characteristics for Heating Load Predictions. *Appl. Energy* **2024**, *359*, 122668. [\[CrossRef\]](#)
35. Roberti, F.; Oberegger, U.F.; Gasparella, A. Calibrating Historic Building Energy Models to Hourly Indoor Air and Surface Temperatures: Methodology and Case Study. *Energy Build.* **2015**, *108*, 236–243. [\[CrossRef\]](#)
36. Ni, Z.; Zhang, C.; Karlsson, M.; Gong, S. A Study of Deep Learning-Based Multi-Horizon Building Energy Forecasting. *Energy Build.* **2024**, *303*, 113810. [\[CrossRef\]](#)
37. Semeraro, S.; Vecchi, F.; Stasi, R.; Berardi, U. Physics-Informed Neural Networks for Predicting Indoor Temperature and Cooling Demand in Historic Buildings. *J. Build. Eng.* **2025**, *115*, 114392. [\[CrossRef\]](#)
38. Hong, T.; Fan, S. Probabilistic Electric Load Forecasting: A Tutorial Review. *Int. J. Forecast.* **2016**, *32*, 914–938. [\[CrossRef\]](#)
39. Chen, Y.; Tan, H. Short-Term Prediction of Electric Demand in Building Sector via Hybrid Support Vector Regression. *Appl. Energy* **2017**, *204*, 1363–1374. [\[CrossRef\]](#)
40. Dudek, G. Short-Term Load Forecasting Using Random Forests. In *Intelligent Systems'2014*; Filev, D., Jablowski, J., Kacprzyk, J., Krawczak, M., Popchev, I., Rutkowski, L., Sgurev, V., Sotirova, E., Szynekarczyk, P., Zadrozny, S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2015; Volume 323, pp. 821–828.
41. Hakkal, S.; Lahcen, A.A. XGBoost to Enhance Learner Performance Prediction. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100254. [\[CrossRef\]](#)
42. Abu-El-Magd, M.A.; Findlay, R.D. A New Approach Using Artificial Neural Network and Time Series Models for Short Term Load Forecasting. In *Proceedings of the CCECE 2003—Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436)*; IEEE: Montreal, QC, Canada, 2003; Volume 3, pp. 1723–1726.
43. Somu, N.; Raman M R, G.; Ramamritham, K. A Hybrid Model for Building Energy Consumption Forecasting Using Long Short Term Memory Networks. *Appl. Energy* **2020**, *261*, 114131. [\[CrossRef\]](#)
44. Dong, B.; Li, Z.; Rahman, S.M.M.; Vega, R. A Hybrid Model Approach for Forecasting Future Residential Electricity Consumption. *Energy Build.* **2016**, *117*, 341–351. [\[CrossRef\]](#)
45. Chen, Y.; Guo, M.; Chen, Z.; Chen, Z.; Ji, Y. Physical Energy and Data-Driven Models in Building Energy Prediction: A Review. *Energy Rep.* **2022**, *8*, 2656–2671. [\[CrossRef\]](#)
46. Harish, V.S.K.V.; Kumar, A. A Review on Modeling and Simulation of Building Energy Systems. *Renew. Sustain. Energy Rev.* **2016**, *56*, 1272–1292. [\[CrossRef\]](#)

47. Crawley, D.B.; Lawrie, L.K.; Winkelmann, F.C.; Buhl, W.F.; Huang, Y.J.; Pedersen, C.O.; Strand, R.K.; Liesen, R.J.; Fisher, D.E.; Witte, M.J.; et al. EnergyPlus: Creating a New-Generation Building Energy Simulation Program. *Energy Build.* **2001**, *33*, 319–331. [\[CrossRef\]](#)
48. Trčka, M.; Hensen, J.L.M. Overview of HVAC System Simulation. *Autom. Constr.* **2010**, *19*, 93–99. [\[CrossRef\]](#)
49. Westphal, F.S.; Lamberts, R. Building simulation calibration using sensitivity analysis. In Proceedings of the Ninth International IBPSA Conference, Montréal, QC, Canada, 15–18 August 2005; pp. 1331–1338.
50. Chargui, R.; Sammouda, H.; Farhat, A. Geothermal Heat Pump in Heating Mode: Modeling and Simulation on TRNSYS. *Int. J. Refrig.* **2012**, *35*, 1824–1832. [\[CrossRef\]](#)
51. Tuhus-Dubrow, D.; Krarti, M. Genetic-Algorithm Based Approach to Optimize Building Envelope Design for Residential Buildings. *Build. Environ.* **2010**, *45*, 1574–1581. [\[CrossRef\]](#)
52. Siddharth, V.; Ramakrishna, P.V.; Geetha, T.; Sivasubramaniam, A. Automatic Generation of Energy Conservation Measures in Buildings Using Genetic Algorithms. *Energy Build.* **2011**, *43*, 2718–2726. [\[CrossRef\]](#)
53. Mazzeo, D.; Matera, N.; Cornaro, C.; Oliveti, G.; Romagnoni, P.; De Santoli, L. EnergyPlus, IDA ICE and TRNSYS Predictive Simulation Accuracy for Building Thermal Behaviour Evaluation by Using an Experimental Campaign in Solar Test Boxes with and without a PCM Module. *Energy Build.* **2020**, *212*, 109812. [\[CrossRef\]](#)
54. Pachano, J.E.; Nuevo-Gallardo, C.; Fernández Bandera, C. An Empirical Comparison of a Calibrated White-Box versus Multiple LSTM Black-Box Building Energy Models. *Energy Build.* **2025**, *333*, 115485. [\[CrossRef\]](#)
55. Meng, Q.; Xiong, C.; Mourshed, M.; Wu, M.; Ren, X.; Wang, W.; Li, Y.; Song, H. Change-Point Multivariable Quantile Regression to Explore Effect of Weather Variables on Building Energy Consumption and Estimate Base Temperature Range. *Sustain. Cities Soc.* **2020**, *53*, 101900. [\[CrossRef\]](#)
56. Alwetaishi, M.; Benjeddou, O. Impact of Window to Wall Ratio on Energy Loads in Hot Regions: A Study of Building Energy Performance. *Energies* **2021**, *14*, 1080. [\[CrossRef\]](#)
57. Li, X.; Ying, Y.; Xu, X.; Wang, Y.; Hussain, S.A.; Hong, T.; Wang, W. Identifying Key Determinants for Building Energy Analysis from Urban Building Datasets. *Build. Environ.* **2020**, *181*, 107114. [\[CrossRef\]](#)
58. Laaroussi, Y.; Bahrar, M.; El Mankibi, M.; Draoui, A.; Si-Larbi, A. Occupant Presence and Behavior: A Major Issue for Building Energy Performance Simulation and Assessment. *Sustain. Cities Soc.* **2020**, *63*, 102420. [\[CrossRef\]](#)
59. Chen, S.; Zhang, G.; Xia, X.; Setunge, S.; Shi, L. A Review of Internal and External Influencing Factors on Energy Efficiency Design of Buildings. *Energy Build.* **2020**, *216*, 109944. [\[CrossRef\]](#)
60. Olu-Ajayi, R.; Alaka, H.; Egwim, C.; Grishikashvili, K. Comprehensive Analysis of Influencing Factors on Building Energy Performance and Strategic Insights for Sustainable Development: A Systematic Literature Review. *Sustainability* **2024**, *16*, 5170. [\[CrossRef\]](#)
61. Akhiat, Y.; Manzali, Y.; Chahhou, M.; Zinedine, A. A New Noisy Random Forest Based Method for Feature Selection. *Cybern. Inf. Technol.* **2021**, *21*, 10–28. [\[CrossRef\]](#)
62. Borowski, M.; Zwolińska, K. Prediction of Cooling Energy Consumption in Hotel Building Using Machine Learning Techniques. *Energies* **2020**, *13*, 6226. [\[CrossRef\]](#)
63. Yoshino, H.; Hong, T.; Nord, N. IEA EBC Annex 53: Total Energy Use in Buildings—Analysis and Evaluation Methods. *Energy Build.* **2017**, *152*, 124–136. [\[CrossRef\]](#)
64. Hong, T.; Yan, D.; D’Oca, S.; Chen, C. Ten Questions Concerning Occupant Behavior in Buildings: The Big Picture. *Build. Environ.* **2017**, *114*, 518–530. [\[CrossRef\]](#)
65. Miao, J.; Niu, L. A Survey on Feature Selection. *Procedia Comput. Sci.* **2016**, *91*, 919–926. [\[CrossRef\]](#)
66. Tang, J.; Alelyani, S.; Liu, H. Feature Selection for Classification: A Review. In *Data Classification: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2014; p. 37.
67. Remeseiro, B.; Bolon-Canedo, V. A Review of Feature Selection Methods in Medical Applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [\[CrossRef\]](#)
68. Venkatesh, B.; Anuradha, J. A Review of Feature Selection and Its Methods. *Cybern. Inf. Technol.* **2019**, *19*, 3–26. [\[CrossRef\]](#)
69. Cormack, G.V.; Clarke, C.L.A.; Buettcher, S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*; ACM: Boston, MA, USA, 2009; pp. 758–759.
70. Van Zyl, C.; Ye, X.; Naidoo, R. Harnessing eXplainable Artificial Intelligence for Feature Selection in Time Series Energy Forecasting: A Comparative Analysis of Grad-CAM and SHAP. *Appl. Energy* **2024**, *353*, 122079. [\[CrossRef\]](#)
71. Liu, Z.; Du, J.; Luo, J.; Yuan, Y.; Huang, Q.; He, J. A Stable Feature Selection Method Based on Majority Voting and SHAP for High-Dimensional Metabolomics Data. *Comput. Methods Programs Biomed.* **2025**, *274*, 109170. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [\[CrossRef\]](#)

73. Eriksson, P.; Hermann, C.; Hrabovszky-Horváth, S.; Rodwell, D. EFFESUS Methodology for Assessing the Impacts of Energy-Related Retrofit Measures on Heritage Significance. *Hist. Environ. Policy Pract.* **2014**, *5*, 132–149. [CrossRef]
74. Mazzearella, L. Energy retrofit of historic and existing buildings. The legislative and regulatory point of view. *Energy Build.* **2015**, *95*, 23–31. [CrossRef]
75. Shenyang Municipal Bureau of Natural Resources. Public Notice of the First Survey Catalogue of Historic Buildings in Shenyang. 20 December 2017. Available online: https://zrzyj.shenyang.gov.cn/ztgz/lswhmcbbh/202208/t20220822_4095091.html (accessed on 25 March 2026).
76. Shenyang Municipal Bureau of Natural Resources. Drafting Notes for the Draft “Shenyang Historic and Cultural City Conservation Plan (2021–2035)” (Draft). 28 November 2023. Available online: <https://zrzyj.shenyang.gov.cn/ztgz/lswhmcbbh/202311/P020231130357239327340.docx> (accessed on 25 March 2026).
77. Shenyang Municipal People’s Government. Notice of the Shenyang Municipal People’s Government on Publishing the Fourth Batch Catalogue of Historic Buildings (Shen Zheng Fa [2018] No. 22). 1 June 2018. Available online: https://www.shenyang.gov.cn/zwgk/zcwj/zfwj/szfwj/202112/t20211201_1701672.html (accessed on 25 March 2026).
78. GB 50176–2016; Thermal Design Code for Civil Buildings. Ministry of Housing and Urban-Rural Development of the People’s Republic of China, China Architecture & Building Press: Beijing, China, 2016.
79. Bo, R.; Chang, W.-S.; Yu, Y.; Xu, Y.; Guo, H. Overheating of residential buildings in the severe cold and cold regions of China: The gap between building policy and performance. *Build. Environ.* **2022**, *225*, 109601. [CrossRef]
80. Shenyang Municipal Bureau of Natural Resources. Reply to Proposal No. 205 of the Second Session of the 16th CPPCC Shenyang Committee: Countermeasures for the Adaptive Reuse of Shenyang Historic Building. 20 April 2023. Available online: https://zrzyj.shenyang.gov.cn/zwgkzdgz/fdzdgknr/jyta/szxjyta/202309/t20230908_4524496.html (accessed on 25 March 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.