buildings

MDPI

Article

Minimum Intervention Assessment in Historic Building Conservation: An Entropy-Weighted Intervention Intensity Index and Machine Learning Analysis at Caishen Temple, Xinzhou

Tianxi Lu [1,2,*], Guihua Zu [3], Siti Sarah Binti Herman [2] and Yang Wang [4]

[1] School of Civil Engineering and Architecture, Jiaozuo University, Jiaozuo 454000, China
[2] Faculty of Design and Architecture, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia
[3] School of Arts, Jiaozuo University, Jiaozuo 454000, China
[4] Tianjin Xinfei Digital Technology Co., Ltd., No. 1 Fengze Road, Zhangjiawo Industrial Park, Xiqing District, Tianjin 300380, China
* Correspondence: gs66799@student.upm.edu.my; Tel.: +60-173837539

Abstract

Conservation interventions in historic buildings require quantitative assessment to balance preservation needs and minimal intervention principles. This study presents a quantitative framework for minimum intervention assessment of heritage conservation, based on 133 documented interventions at Caishen Temple in Xinzhou across four periods: 1992, 2006, 2016, and 2017. Four dimensions were defined: Extent of Intervention (*EI*), Reversibility (*RE*), Information Loss (*IL*), and Necessity (*NE*). Expert scoring on a five-point Likert scale yielded high inter-rater reliability (*ICC*: 0.79–0.88). The entropy weight method was then used to derive data-driven weights from the expert-scoring matrix, $IL = 0.3149$, $EI = 0.3025$, $NE = 0.2572$, $RE = 0.1253$, and the Intervention Intensity Index (*III*) was calculated for each intervention. A random forest model was further developed as an exploratory cross-check, with 19 problematic interventions labelled as $y = 1$ and 114 normal interventions as $y = 0$. Target labels were defined through a dual-source procedure combining SSIM- and HSV-based image-similarity assessment for interventions with paired *pre*- and *post*-restoration photographs and explicit textual evidence from conservation records for interventions without paired photographs. Feature importance and SHAP analyses indicated that reversibility and necessity had the greatest discriminative importance for distinguishing problematic from normal interventions in this dataset, whereas the entropy-based ranking assigned higher weights to information loss and extent of intervention. This descriptive contrast, based on only four dimensions, highlights the distinction between data dispersion and discriminative capacity. Spatiotemporal analysis indicated that roof and rafter components exhibited the highest intervention intensity, while 1992 interventions contained the largest proportion of problematic practices. This study combines entropy-derived weighting with a machine-learning-based exploratory cross-check, providing a transparent, case-based framework for evidence-based conservation decision-making at historic heritage sites.

Keywords: minimum intervention; heritage conservation; intervention intensity index; entropy weight method; random forest; machine learning

Check for updates

Academic Editor: Cinzia Buratti

Received: 15 June 2026
Revised: 17 July 2026
Accepted: 1 August 2026
Published: 5 August 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) license.

1. Introduction

The conservation of historic architectural heritage represents a critical balance between safeguarding cultural authenticity and implementing necessary restorative measures, as

guided by the minimum intervention principle [1,2]. This principle, widely endorsed in international charters, calls for interventions to be limited to what is strictly necessary, preserving as much original fabric as possible. However, translating this qualitative principle into practice remains challenging due to the lack of systematic, quantitative tools to assess intervention intensity [3].

In recent years, computational methods have been increasingly adopted in heritage research. Machine learning techniques have shown promise in automated damage detection, such as identifying pathologies of shedthin tiles (thin clay tiles used as a secondary roofing layer in traditional Chinese architecture [4]), virtual restoration [5], detecting surface anomalies on heritage roofs [6], assessing material deterioration in masonry towers [7], and detecting chromatic decay in tropical heritage buildings [8]. These studies focus on recognizing physical degradation rather than evaluating the appropriateness of conservation actions. Simultaneously, material science investigations [9,10] provide essential knowledge on degradation mechanisms but do not address intervention decision-making.

Beyond pathology detection, several researchers have developed multi-criteria frameworks to support conservation decisions. Sánchez-Calvillo et al. compared multiple seismic vulnerability assessment methods (GNDT, Vicente, SVIVA, SAVVAS) for adobe buildings, demonstrating how parameter-based indices can quantify structural risk [11]. Fiore et al. used analytic hierarchy process (AHP) to evaluate integrated interventions on historic buildings, considering preservation, seismic safety, and energy efficiency [12]. Maselli et al. proposed an AHP-based model (H-MCDM) with seven indicators for adaptive reuse selection, aligning with sustainability goals [13]. Zhang combined AHP and fuzzy comprehensive evaluation to determine minimum intervention levels for 14 historical buildings along Beijing's central axis [14]. Liu et al. applied entropy weight method and GRA-TOPSIS for post-occupancy evaluation of railway heritage, using cosine similarity to neutralize individual preferences [15]. Bertolin and Loli developed a zero-emission refurbishment tool that links intervention levels with life-cycle emissions [16]. Cardinale et al. compared traditional and modern seismic interventions, highlighting the lower environmental impact of traditional techniques [17].

Despite these advances, existing frameworks primarily target structural safety, energy efficiency, or adaptive reuse potential. Few directly assess the intensity of conservation interventions themselves with respect to the minimum intervention principle. Moreover, most studies rely on subjective weighting (e.g., AHP) that may embed expert biases, and they rarely compare their weighting schemes with machine-learning-based discriminative analyses.

To address these gaps, this study develops a quantitative framework integrating the entropy weight method and random forest analysis. Four core dimensions are defined: Extent of Intervention (*EI*), Reversibility (*RE*), Information Loss (*IL*), and Necessity (*NE*). Expert scoring and intraclass correlation coefficient (*ICC*) analysis ensure reliability [18]. The entropy weight method—used objectively in heritage evaluation—determines dimension weights, which are then combined into an Intervention Intensity Index (*III*) [15]. A random forest model was used to cross-check the weight ranking and identify key predictors of inappropriate interventions, drawing on machine-learning approaches demonstrated in previous heritage-detection studies [4,5]. Spatiotemporal analysis of 133 documented interventions at Caishen Temple reveals patterns of over-intervention [19].

Located in Xinzhou, China, Caishen Temple is a vernacular heritage site that embodies traditional northern Chinese architecture (Figure 1). Its conservation records document 133 intervention actions carried out or proposed across four periods: 1992, 2006, 2016, and 2017. These interventions cover a wide range of building components, including roof structures, timber frames, masonry walls, decorative elements, platforms, and environmental features.

This well-documented dataset provides a suitable basis for systematically analysing intervention patterns, assessing compliance with the minimum-intervention principle, and identifying potential risks in conservation practice.

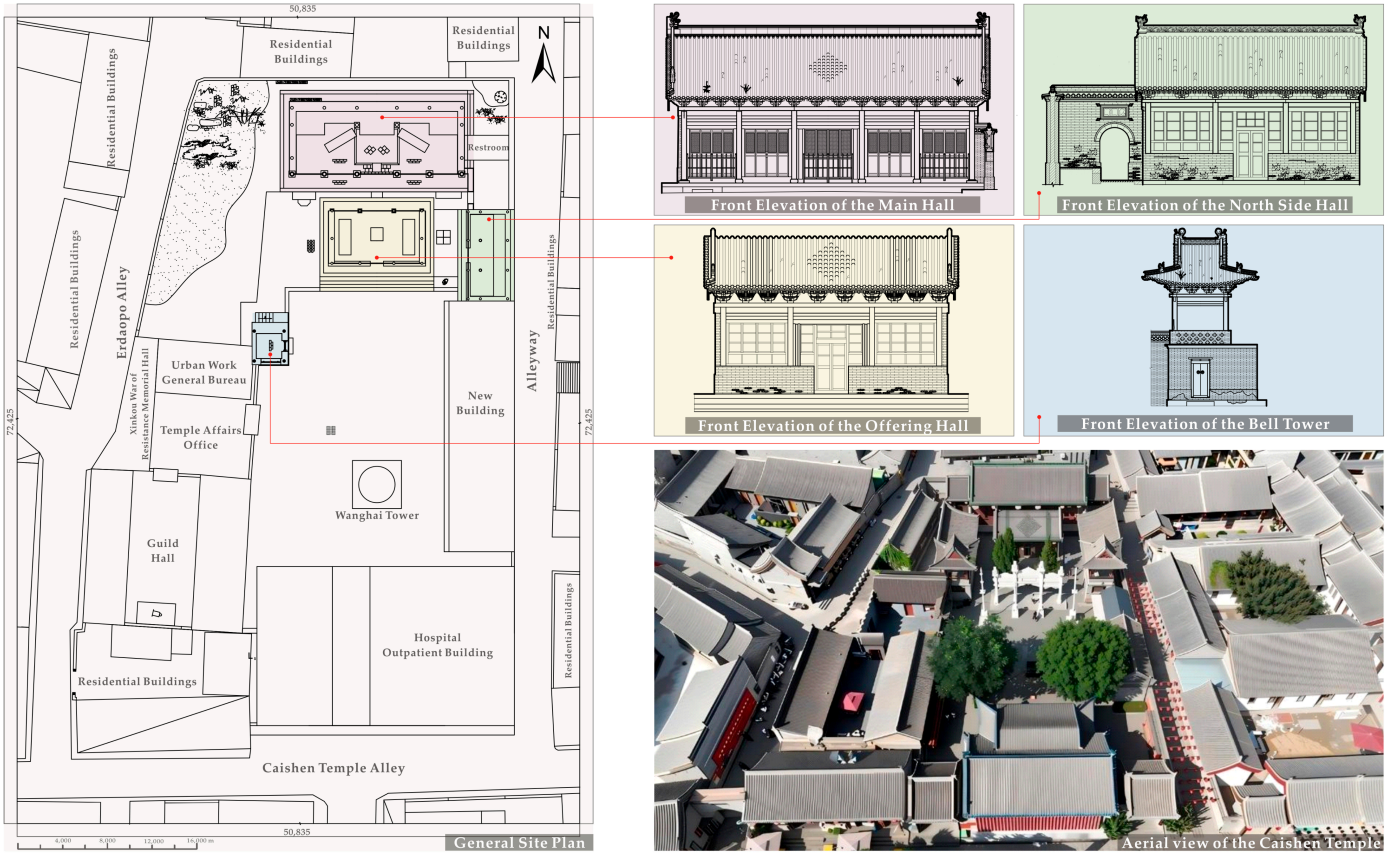


Figure 1. Architectural Overview of Caishen Temple in Xinzhou.

Despite the availability of detailed intervention records, existing research on Caishen Temple remains largely descriptive, lacking a standardized quantitative framework to evaluate intervention intensity and its underlying drivers. Most studies focus on architectural morphology or material analysis, with limited attention paid to the systematic assessment of conservation behavior. This gap hinders evidence-based conservation planning and long-term heritage management.

To address this issue, this study develops a quantitative assessment framework integrating the entropy weight method and random forest analysis. Four core evaluation dimensions are defined: Extent of Intervention (*EI*), Reversibility (*RE*), Information Loss (*IL*), and Necessity (*NE*). Expert scoring and intraclass correlation coefficient (*ICC*) analysis are conducted to ensure evaluation reliability. The entropy weight method is used to calculate entropy-based weights, which are then integrated into the Intervention Intensity Index (*III*) for quantitative evaluation. A random forest model was used to cross-check the weight ranking and identify key predictors of interventions flagged as problematic through the dual-source labelling procedure described in Section 2.4. Spatiotemporal patterns of intervention intensity are analyzed to highlight high-risk components and periods.

This study addresses the following research question: Can the invasiveness of conservation interventions be quantified in a way that distinguishes appropriate repairs from over-intervention, and which dimensions of intervention are most predictive of inappropriateness? The main contributions of this study are as follows: (1) shifting quantitative heritage assessment from the evaluation of building damage or structural performance to the assessment of conservation intervention intensity and appropriateness; (2) establishing

a standardized four-dimensional framework—Extent of Intervention, Reversibility, Information Loss, and Necessity—for operationalising the minimum-intervention principle in heritage conservation; (3) applying entropy weighting, random forest classification, and SHAP interpretation as complementary analytical tools to examine how intervention-intensity scores relate to the identification of problematic conservation actions; and (4) providing empirical evidence from Caishen Temple, showing that roof and rafter components exhibit the highest intervention intensity and that interventions conducted in 1992 contained the largest proportion of problematic practices. The proposed framework provides a transparent and reproducible case-based assessment procedure for supporting evidence-based conservation decision-making and sustainable heritage management.

2. Materials and Methods

The overall technical workflow of this study comprises four main stages, as illustrated in Figure 2. The methodological process begins with the systematic construction of a conservation intervention database, followed by expert evaluation and reliability testing, data-derived weighting via the entropy weight method, and finally, an exploratory cross-check using a machine-learning approach.

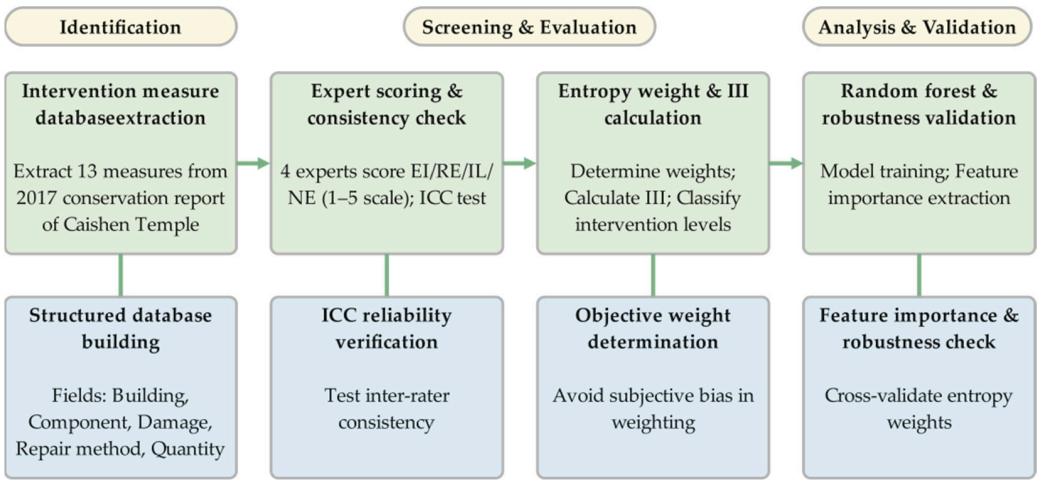


Figure 2. Technical Workflow of Minimum Intervention Assessment.

In the first stage, a comprehensive intervention database was established based on the official conservation archives of Caishen Temple in Xinzhou. A total of 133 documented conservation measures were extracted from the 2017 conservation design report, covering a wide range of restoration and repair activities implemented across the site. A structured database was built with standardized fields, including building name, architectural component, damage description, repair method, and intervention quantity, ensuring data consistency and traceability.

In the second stage, expert evaluation was conducted to assess the performance of each intervention. Four experts with professional backgrounds in historic building restoration, heritage conservation, and materials science were invited to independently evaluate each intervention using a 5-point Likert scale [20]. Four core evaluation dimensions were defined: Extent of Intervention (*EI*), Reversibility (*RE*), Information Loss (*IL*), and Necessity (*NE*). The intraclass correlation coefficient (*ICC*) was calculated to verify inter-rater consistency and ensure the reliability of expert scoring.

In the third stage, the entropy weight method was employed to determine data-derived weights for the four evaluation dimensions. The weighting calculation was based solely on the 133 × 4 expert scoring matrix, deriving weights directly from the dispersion of the scoring data without requiring subjective prior assumptions. The Intervention Inten-

sity Index (*III*) for each intervention was then computed via linear weighted summation. Threshold values were further defined to classify interventions into low, medium, and high intervention levels.

In the fourth stage, a random forest model was introduced as an exploratory cross-check to compare the entropy-based weight ranking with a discriminative importance ranking derived from the same data. The model used the inappropriate alteration labels ($y = 1$ for problematic, $y = 0$ for normal) defined through the dual-source labelling procedure described below as the target variable. Feature importance analysis and cross-validation were performed to assess model stability and compare the discriminative importance ranking with the entropy-based weight ranking. The following subsections describe each stage in detail.

2.1. Construction of the Intervention Measure Database

This study established a structured intervention database based on the official conservation archives of Caishen Temple in Xinzhou. A total of 133 documented interventions were systematically compiled, covering four key periods: 1992, 2006, 2016, and 2017. Among these, 103 measures (77.4%) were proposed in the 2017 conservation design report, while the remaining 30 measures represent documented interventions from earlier periods (1992, 2006, and 2016). The interventions were classified according to the traditional Chinese architectural component system, including roofs, timber structures, masonry walls, platforms, and auxiliary facilities, to ensure alignment with Chinese heritage conservation conventions.

A standardized structured database was established, with core fields including ID, building name, architectural component, damage description, repair method, and intervention quantity (Table 1). Each intervention was assigned a unique ID for unified management and subsequent analysis. The database ensures the traceability and consistency of all intervention records, providing a reliable data foundation for quantitative assessment.

Table 1. Example of intervention measures from the conservation design report.

ID	Building	Component	Image	Damage Description	Repair Method	Quantity	Year of Intervention
1	Main Hall	Front roof slope		30% of tiles frost-fractured, 20 eaves-end tiles missing each	Remove and re-lay roof, supplement 70 m^2 of pan-and-roll tiles, 32 eaves-end tiles, 32 drip tiles	70 m^2, 32 + 32 pieces	2006
21	Main Hall	Gable wall (door/window holes, south)		Artificial door/window holes opened, lower part salt-efflorescent	Remove and seal the openings, replace salt-efflorescent bricks	3.6 m^2	1992
33	Offering Hall	Roof		30% of tiles frost-fractured, 30% of eaves-end tiles missing front and rear	Remove and re-lay roof 78.3 m^2, supplement 14.36 m^2 of pan-and-roll tiles	78.3 m^2	2017

Table 1. *Cont.*

ID	Building	Component	Image	Damage Description	Repair Method	Quantity	Year of Intervention
47	Offering Hall	Stylobate (cement, front steps)		Front platform cement-plastered during recent repairs in 2006	Remove cement, lay brick wall-base area	2.36 m^2	1992
64	North Side Hall	Joinery (formal alteration, decorative finishes)		Original decorative finishes lost, changed to eaves-column installation	Remove modern doors/windows, restore original gold-column installation according to mortise marks	8.3 m^2	2017
84	North Side Hall	Front eaves rafter		25 front eaves rafters decayed and cracked	Replace 52 eaves rafters	107.38 m^2	2006
97	Bell Tower	Front roof slope		10% of tiles frost-fractured, 3 eaves-end tiles missing each	Remove and re-lay roof 35 m^2, supplement 3 m^2 of pan-and-roll tiles	35 m^2	2017
116	Bell Tower	Decorative balustrade (brick, second-floor)		Second-floor decorative balustrade arbitrarily built with machine-made bricks during 2006 repair	Dismantle and rebuild with blue bricks	2.7 m^2	2006
130	Environment	Courtyard surface (front, cobble → red brick)		Original cobble courtyard removed, replaced with red brick paving	Restore cobble courtyard surface	1500 m^2	1992
132	Environment	Ancient well		Cover stone slabs collapsed, previously used for wastewater discharge	Emergency repair of cover stone slabs (already done in 2016)	–	2016

The distribution of interventions is presented in Figure 3, which includes two subplots. Figure 3a shows the distribution of interventions across different architectural components. In traditional Chinese timber-framed architecture, roofs and timber structures are the most vulnerable and frequently repaired components, consistent with the long-term conservation

records of Caishen Temple. Figure 3b illustrates the temporal distribution of interventions. Conservation activities were concentrated in 2017, with the 1992 and 2006 interventions representing earlier ad-hoc repairs and the 2016 measures being emergency actions. The systematic database provides a solid foundation for the subsequent multi-dimensional assessment of intervention appropriateness.

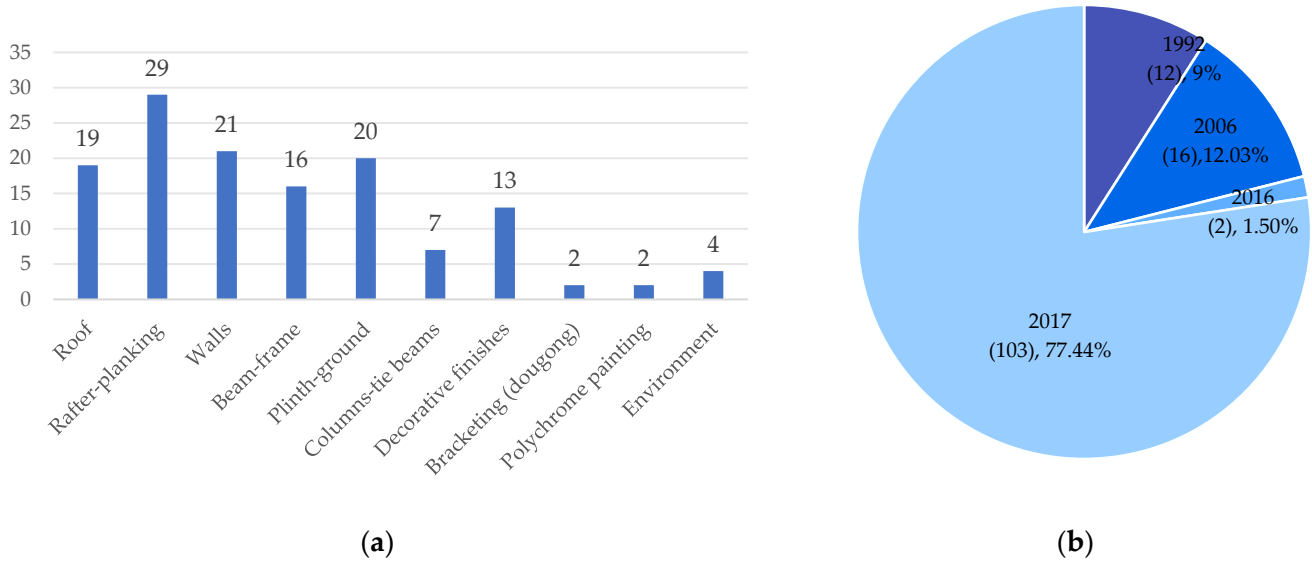


Figure 3. Distribution of intervention measures: (**a**) component category; (**b**) historical period.

2.2. Expert Scoring and Reliability Assessment

Four experts with professional backgrounds in historic building restoration, heritage conservation, materials science, and on-site survey of Caishen Temple were invited to independently score each intervention. The Intervention Intensity Index (*III*) is characterized by four dimensions: Extent of Intervention (*EI*), Reversibility (*RE*), Information Loss (*IL*), and Necessity (*NE*). All dimensions were scored on a 5-point Likert scale (1–5). For consistency with the risk-oriented scoring system, higher scores in all four dimensions indicate higher intervention intensity or higher conservation risk. In the NE dimension, a higher score therefore indicates lower intervention necessity rather than stronger necessity. The definitions and scoring criteria are summarized in Table 2. The scoring process included a 30-min training session and a pilot scoring of 10 pre-selected measures, followed by independent scoring of all 133 measures. To avoid scoring drift, the work was divided into three batches with at least 24 h between batches.

Table 2. *III* dimension scoring criteria.

Score	Extent of Intervention (*EI*)	Reversibility (*RE*)	Information Loss (*IL*)	Necessity (*NE*)
1	No replacement (cleaning, consolidation)	Fully reversible (temporary support)	No loss (recording, dusting)	Absolutely necessary (imminent collapse)
2	Partial replacement (<30%)	Conditionally reversible (lime mortar)	Minor loss (surface cleaning)	Necessary (moderate decay)
3	Moderate replacement (30–70%)	Moderately reversible	Moderate loss (partial replacement)	Moderate (can be deferred)
4	Extensive replacement (>70%)	Poorly reversible (cement mortar)	Substantial loss (extensive replacement)	Optional (minor damage)
5	Complete replacement or formal alteration	Irreversible (cement hardening)	Severe loss (demolition, covering)	Unnecessary (functional change)

The intraclass correlation coefficient (*ICC*) was calculated using a two-way random effects model ($ICC(2, k)$) to assess inter-rater reliability. *ICC* values for each dimension were computed, and dimensions with *ICC* below 0.75 would trigger a re-examination of measures with large disagreement (standard deviation > 1.0) until consensus was reached [17]. The arithmetic mean of the four experts' scores was taken as the final dimension score for each measure (rounded to one decimal place).

2.3. Weight Determination via the Entropy Weight Method

The entropy weight method, a data-derived weighting approach widely used in multi-criteria decision analysis, was applied to the 133 × 4 expert-mean scoring matrix to determine weights for the four dimensions. The procedure followed standard steps: range normalisation, entropy calculation, and weight normalisation. The Intervention Intensity Index (*III*) for each measure was then calculated as the linear weighted sum:

$$III_i = w_{DI}DI_i + w_{RE}RE_i + w_{IL}IL_i + w_{NE}NE_i, \quad (1)$$

Because all four dimensions are scored on a 1–5 scale and are positively oriented (higher score = higher intervention intensity), the resulting III_i also ranges from 1 to 5. Based on the distribution of III values and expert recommendations, intervention levels were defined as low ($III < 2.5$), medium ($2.5 \leq III < 3.5$) and high ($III \geq 3.5$). The appropriateness of these thresholds was subsequently validated using cluster analysis.

2.4. Cross-Validation Using Random Forest

To compare the entropy-based weight ranking with the discriminative importance ranking, a random forest classifier was applied and evaluated using cross-validation [21,22]. The features (X) were the 133 × 4 expert-mean scoring matrix (*EI*, *RE*, *IL*, *NE*). To address the potential circularity concern that both the predictor variables and the target labels were derived from the same expert-scoring framework, this study defined the 'inappropriate alteration' label ($y = 1$) using a separate data-driven labelling procedure. For each intervention with available pre- and post-restoration photographic records, the visual fidelity of the restoration was quantified using two complementary metrics: (1) the structural similarity index (SSIM) between paired images and (2) colour difference in HSV colour space [23]. For interventions with available paired photographic records, those falling below the 25th percentile of the similarity distribution were labelled as problematic ($y = 1$), while the remaining photographed interventions were labelled as normal ($y = 0$). For interventions without paired photographic records, labels were assigned on the basis of explicit textual evidence in the conservation report indicating irreversible formal alterations or the use of incompatible materials. This dual-source labelling strategy reduces direct dependence on the expert-scoring framework used for the predictor dimensions, although statistical association between the target labels and the scored dimensions remains possible.

To address class imbalance, the random forest was configured with $class_{weight} = balanced$. The classifier was implemented using the scikit-learn library in Python. Given the modest sample size (133 cases) and the small number of positive cases (19), the main hyperparameters were set with a shallow $max_{depth} = 3$ to prevent overfitting, together with $n_{estimators} = 100$, $criterion = 'gini'$, and $random_{state} = 42$. Feature importance was first obtained from the random forest model using impurity-based importance and was further interpreted using SHAP values to improve model transparency. Five-fold stratified cross-validation was applied, and the average feature importance across the five folds was normalised to sum to unity. The resulting importance ranking was compared with the entropy-based weight ranking using Spearman's rank correlation coefficient.

2.5. Statistical Analysis and Software

All statistical analyses and machine learning procedures were performed using Python 3.12. Key libraries included: pandas (version 1.5.3) and numpy (1.23.5) for data manipulation; scikit-learn (1.2.2) for random forest implementation, cross-validation, and feature importance calculation; and SciPy (1.10.0) for *ICC* computation and Spearman's rank correlation. The entropy weight method was implemented via a custom script following the standard algorithm. Visualisations (heatmaps, boxplots, etc.) were generated using matplotlib (3.6.2) and seaborn (0.12.2). All analyses were performed on a workstation with an Intel Core i7-12700 CPU and 32 GB RAM. The complete code and anonymised dataset are available from the corresponding author upon reasonable request.

3. Results

3.1. Reliability of Expert Scoring and Entropy-Based Weights

The intraclass correlation coefficient (*ICC*) was calculated to assess the consistency of the four independent experts' scores across all 133 interventions. For the four evaluation dimensions (Extent of Intervention, *EI*; Reversibility, *RE*; Information Loss, *IL*; Necessity, *NE*)—the $ICC(2, k)$ values were 0.86, 0.79, 0.88, and 0.82, respectively. All values exceeded the commonly accepted threshold of 0.75 for good reliability, indicating a high level of inter-rater agreement. Therefore, the arithmetic mean of the four experts' scores was used as the final dimension score for each intervention.

The entropy weight method was then applied to the 133×4 expert-mean scoring matrix to determine data-derived weights for the four dimensions. The calculated weights are presented in Table 3. Information Loss received the highest weight (0.3149), followed by Extent of Intervention (0.3025), Necessity (0.2572), and Reversibility (0.1253). This ranking indicates that, from a data dispersion perspective, *IL* and *EI* are the most discriminating dimensions for differentiating intervention intensity, while *RE* contributes the least to the variance among the interventions.

Table 3. Data-derived weights of the four evaluation dimensions derived from the entropy weight method.

Dimension	Symbol	Weight
Information Loss	*IL*	0.3149
Extent of Intervention	*EI*	0.3025
Necessity	*NE*	0.2572
Reversibility	*RE*	0.1253

3.2. Spatiotemporal Patterns of Intervention Intensity

Using the entropy-based weights and the dimension scores for each intervention, the Intervention Intensity Index (*III*) was calculated for all 133 measures. The *III* values ranged from 1 to 5 across the 133 measures, with a mean of 2.91 ($SD = 0.89$). The minimum value (1.00) corresponds to a non-invasive mural protection measure (ID63), while the maximum (5.00) corresponds to several inappropriate interventions such as artificial openings and cement hardening (e.g., ID21, ID47, ID51). A k-means clustering analysis performed on the *III* values confirmed three distinct groups whose centroids (2.13, 3.02, and 4.05) aligned closely with the threshold boundaries (low: $III < 2.5$; medium: $2.5 \leq III < 3.5$; high: $III \geq 3.5$), supporting the validity of the three-level classification. The spatial distribution of intervention intensity across building typologies and component categories is visualised as a heatmap in Figure 4.

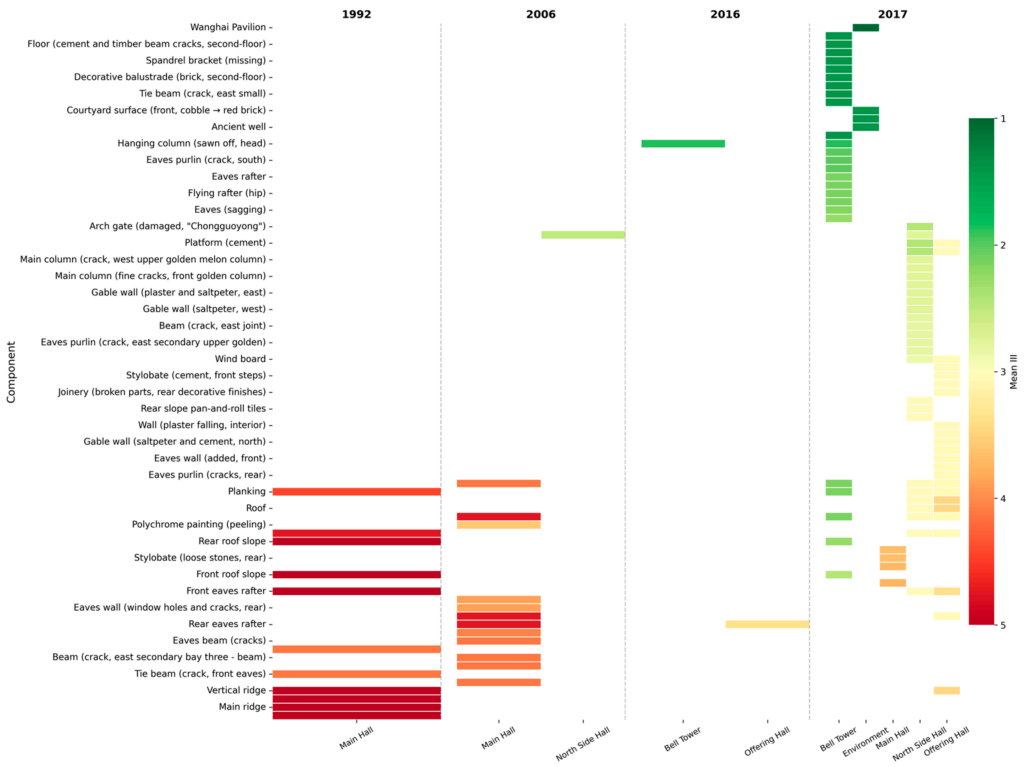


Figure 4. Heatmap of mean III values by building component and historical period.

The heatmap shows that interventions in 1992 and 2006 were predominantly high-intensity and concentrated on non-structural elements such as walls and floors. In contrast, the 2017 interventions covered a wider range of core structural components (beams, columns, rafters), with a mix of low, medium, and high intensities. A high-intensity cluster was observed at the intersection of the North Side Hall and decorative/wall components.

The temporal evolution of intervention intensity is shown in Figure 5. Interventions conducted in 1992 had the highest median *III* (close to 5), followed by 2006 (*median* $\approx$ 4), while 2016 and 2017 showed much lower medians (2.61 and 2.79, respectively). A one-way ANOVA revealed significant differences among the four periods ($F(3,\ 129) = 18.64$, $p < 0.001$), with post-hoc Tukey tests confirming that the 1992 group differed significantly from all later periods ($p < 0.01$).

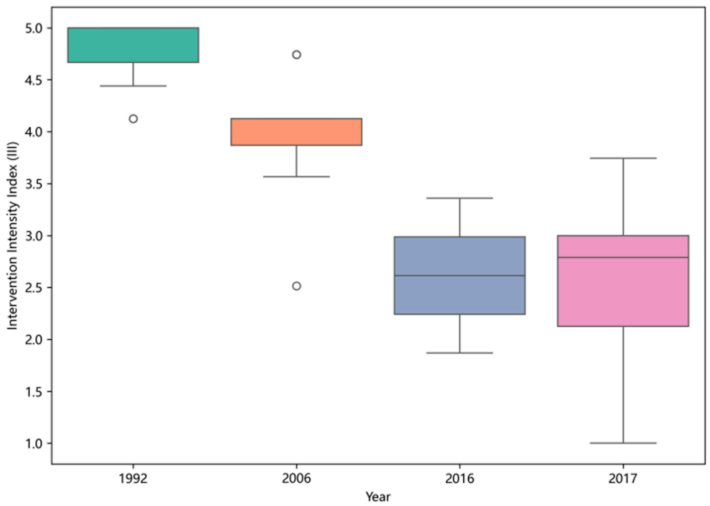
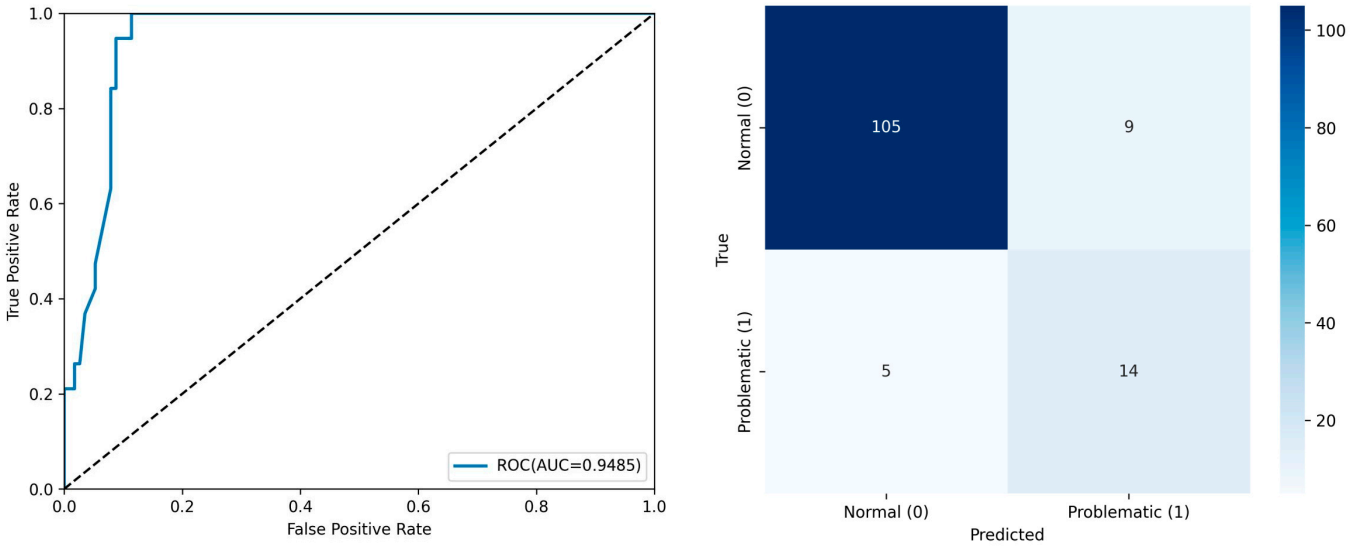


Figure 5. Boxplot of III values across four historical periods, 1992–2017.

3.3. Random Forest Performance and Feature Importance

The random forest classifier was trained and validated using five-fold stratified cross-validation on the 133 × 4 expert-mean scoring matrix, with the 'inappropriate alteration' label (19 positive, 114 negative; defined through the dual-source labelling procedure described in Section 2.4) as the target variable. Over 50 repeats of five-fold stratified cross-validation, the model achieved a mean AUC of 0.9485 (95% CI: [0.9361, 0.9595], SD = 0.0055), indicating consistent discriminative performance (Figure 6). However, given the modest sample size (133 total, 19 positive cases) and the strong collinearity among predictors (*EI*–*IL*: 0.91, *RE*–*IL*: 0.88, *RE*–*NE*: 0.73), the AUC should be interpreted cautiously. Impurity-based importance is known to be biased toward correlated features [21,22]. The confusion matrix from a representative run shows 105 TN, 9 FP, 5 FN, and 14 TP. Given the imbalanced class distribution (114 vs. 19), *sensitivity* = 14/19 (73.7%) and *specificity* = 105/114 (92.1%) provide a more meaningful assessment than accuracy alone.



(**a**) (**b**)

Figure 6. Classification performance of the random forest model: (**a**) receiver operating characteristic (ROC) curve; (**b**) confusion matrix.

Two complementary importance measures were computed. Impurity-based importance (mean over 50 CV repeats) ranked $RE(0.403) > NE(0.370) > IL(0.185) > EI(0.042)$. Permutation importance (30 permutations, $scoring = roc_auc$) confirmed the same ranking: $RE(0.054) > NE(0.034) > IL(0.003) > EI(0.002)$. Across 50 CV repeats, RE ranked first in 70% of runs and NE in 30%, indicating moderate rank stability. This ranking differs notably from the entropy-based weight ranking ($IL > EI > NE > RE$) reported in Table 3. Spearman's rank correlation coefficient ρ was -0.8000 ($p = 0.2000$, $n = 4$), indicating that the two methods produce opposite rankings of the four dimensions. However, because the comparison involves only four dimensions, this coefficient is presented descriptively and no inferential claim about statistical significance is warranted. The divergence suggests that entropy weights capture overall data dispersion, whereas random forest importance identifies dimensions most critical for distinguishing problematic interventions.

To provide a clear quantitative summary supporting the following discussion, Table 4 presents the mean *III*, median *III* and high-intensity proportion ($III \geq 3.5$) for each historical period and each component category. The period-wise statistics confirm that interventions from 1992 exhibit the highest mean *III* (4.79) and a high-intensity proportion of 100%, whereas the 2017 newly proposed measures have a mean *III* of 2.53 and a high-intensity

proportion of only 4.9% (5 out of 103). Among component categories, roof (mean 3.65), walls (mean 3.10), and beam-frame (mean 2.98) show the highest mean III values. In contrast, columns–tie beams (mean 2.39, 0% high) and environment (mean 1.29, 0% high) exhibit the lowest intervention intensity.

Table 4. Summary of intervention intensity by period and component category.

Category	Sub-Category	No. of Measures	Mean III	Median III	High-Intensity Proportion (III $\geq$ 3.5)
Period	1992	12	4.79	5.00	100% (12/12)
	2006	16	4.03	4.13	93.8% (15/16)
	2016	2	2.61	2.61	0% (0/2)
	2017	103	2.53	2.79	4.9% (5/103)
Component category	Roof	19	3.65	3.44	36.8% (7/19)
	Rafter-planking	29	3.22	3.05	24.1% (7/29)
	Walls	21	3.10	3.05	33.3% (7/21)
	Beam-frame	16	2.98	2.79	31.2% (5/16)
	Columns–tie beams	7	2.39	2.74	0% (0/7)
	Plinth-ground	20	2.54	2.72	20.0% (4/20)
	Decorative finishes	13	2.11	2.43	0% (0/13)
	Bracketing (dougong)	2	2.75	2.75	50.0% (1/2)
	Polychrome painting	2	3.28	3.28	50.0% (1/2)
	Environment	4	1.29	1.38	0% (0/4)

To further clarify the statistical relationship between the target label and the input dimensions—a key concern regarding potential circularity—Table 5 compares the mean scores of Problem = 1 (n = 19) and Problem = 0 (n = 114) groups across all four dimensions and the III. All four dimensions are significantly higher in the Problem = 1 group (Mann–Whitney U test, all $p < 0.001$), with large effect sizes (Cohen's $d > 1.4$). The effect size ranking mirrors the random forest feature importance ranking (RE > NE > IL > EI), indicating that the dimensions most effective at discriminating problematic from normal interventions also exhibit the largest between-group separation. This association is expected, as interventions scoring high on all dimensions are inherently more likely to be categorized as problematic; the table confirms the internal consistency of the labelling procedure rather than providing independent evidence. The random forest should therefore be interpreted as an exploratory cross-check of discriminative capacity rather than a causal or independent validation.

Table 5. Comparison of dimension scores between Problem = 1 and Problem = 0 groups.

Dimension	Problem = 1 (n = 19)	Problem = 0 (n = 114)	Difference	p-Value	Cohen d
EI	4.32 $\pm$ 0.80	2.86 $\pm$ 1.15	+1.46	<0.001	1.47
RE	4.79 $\pm$ 0.69	2.97 $\pm$ 0.79	+1.82	<0.001	2.44
IL	4.42 $\pm$ 0.75	2.64 $\pm$ 1.02	+1.78	<0.001	1.99
NE	4.11 $\pm$ 0.64	2.35 $\pm$ 0.81	+1.75	<0.001	2.41
III	4.35 $\pm$ 0.63	2.67 $\pm$ 0.85	+1.68	<0.001	2.25

To assess the stability of the feature importance estimates—as requested by the reviewer—bootstrap resampling with 500 iterations was performed. Table 6 reports the 95% confidence intervals for impurity-based importance. RE shows the widest interval ([0.210, 0.575]) but consistently ranks first, reflecting uncertainty in magnitude rather than rank. EI has the largest coefficient of variation (0.57), indicating the least stable importance estimate.

Table 7 reports permutation importance, which is more robust to collinearity [21,22]. The confidence intervals for EI and IL cross zero, indicating that these dimensions do not significantly contribute to predictive performance beyond the four correlated dimensions. In contrast, RE and NE have intervals entirely above zero, confirming them as genuine discriminative features. Across 500 bootstrap samples, RE ranked first in 66% of runs and NE in 28%, with IL and EI consistently ranking third (67%) and fourth (99%), respectively, indicating moderate rank stability overall.

Table 6. Bootstrap 95% confidence intervals for impurity-based feature importance (500 resamples).

Feature	Mean	2.5% CI	97.5% CI	CV
EI	0.0547	0.0170	0.1320	0.57
RE	0.4097	0.2101	0.5750	0.25
IL	0.2230	0.1339	0.3738	0.28
NE	0.3126	0.1370	0.5182	0.33

Method: Bootstrap resampling (n = 133, with replacement), 500 iterations. On each resample, RandomForestClassifier (n_estimators = 100, max_depth = 3) was fitted and impurity-based feature importances were extracted. CI = percentile method (2.5th, 97.5th).

Table 7. Bootstrap 95% confidence intervals for permutation importance (500 resamples).

Feature	Mean	2.5% CI	97.5% CI
EI	0.0038	−0.0017	0.0164
RE	0.0764	0.0057	0.1961
IL	0.0091	−0.0015	0.0499
NE	0.0616	0.0026	0.1905

Method: Bootstrap resampling (n = 133, with replacement), 500 iterations. On each resample, permutation importance with 10 repeats (scoring = roc_auc) was computed. CI = percentile method (2.5th, 97.5th). Note: *EI* and *IL* confidence intervals cross zero, indicating these features do not significantly contribute to predictive performance beyond the four correlated dimensions.

To evaluate the robustness of the AUC against overfitting—particularly given the small positive class (19 cases) and strong feature collinearity—the model was trained and evaluated over 50 repeats of five-fold stratified cross-validation. Table 8 shows the mean AUC of 0.9485 (95% CI: [0.9361, 0.9595]), indicating limited repeated-seed variability within this dataset.

Table 8. AUC from repeated 5-fold stratified cross-validation (50 repeats).

Metric	Mean	2.5% CI	97.5% CI	SD
AUC	0.9485	0.9361	0.9595	0.0055

Method: 5-fold stratified cross-validation, repeated 50 times with different random seeds. AUC computed as the mean across all repeats. CI = percentile method.

3.4. SHAP Interpretation of Problematic Interventions

To further interpret the model predictions for problematic interventions, SHAP analysis was conducted on the trained random forest model. Figure 7 (Global SHAP bar plot of mean absolute SHAP values) confirms the same ranking as the random forest importance: $RE > NE > IL > EI$, with *RE*'s contribution approximately ten times that of *EI*. This suggests that reversibility and necessity are strong predictors of inappropriate interventions, rather than definitive causal determinants.

Figure 8 provides detailed insight into how feature values affect predictions. For *RE*, low values (blue, good reversibility) are concentrated on the negative SHAP side (reducing risk), while high values (red, poor reversibility) appear on the positive side. For *NE*, low scores (blue, high necessity) reduce risk, whereas high scores (red, low necessity) increase risk-unnecessary interventions strongly associated with inappropriateness. *IL* shows that

high information loss (red) drives positive SHAP values, while *EI* points cluster near zero, confirming its minimal impact.

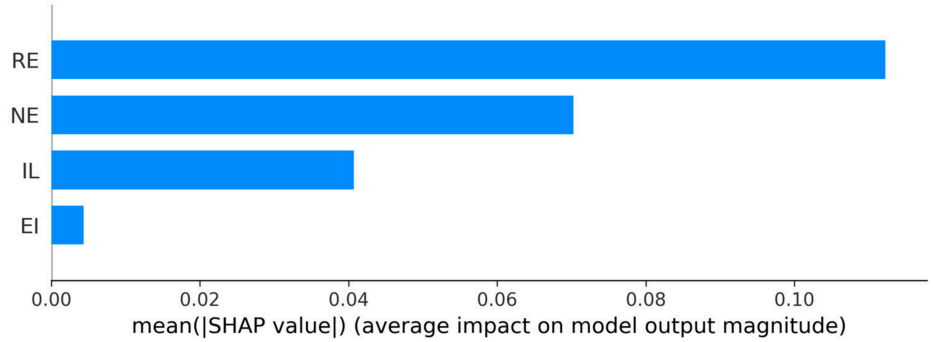


Figure 7. Global SHAP bar plot of mean absolute SHAP values.

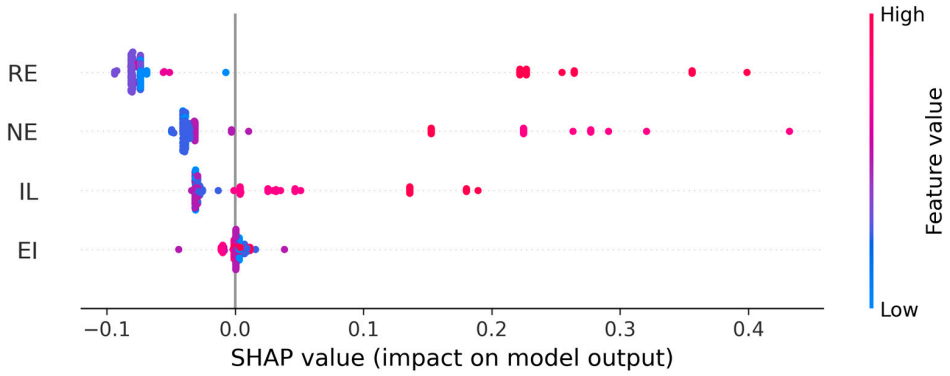


Figure 8. SHAP beeswarm plot for all input features.

The cumulative contribution of features is quantified in Figure 9. The first three features (*RE*, *NE*, *IL*) together account for over 90% of the total predictive contribution, suggesting that *EI* could be omitted in a simplified assessment framework without substantial loss of accuracy.

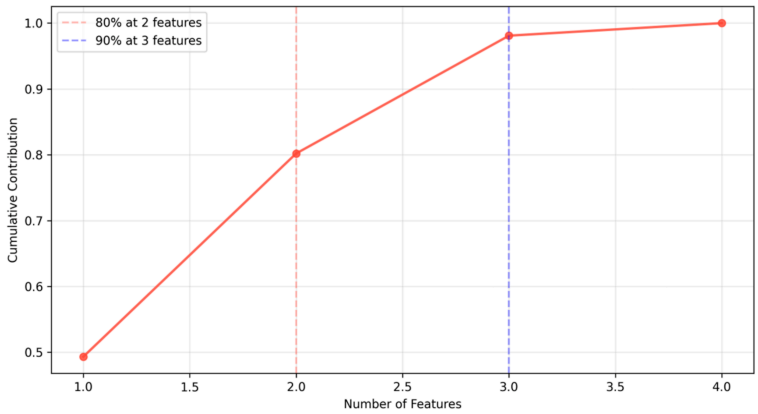


Figure 9. Feature cumulative contribution curve based on SHAP values.

Figure 10 reveals strong positive correlations between *RE* and *IL* (0.77), *RE* and *NE* (0.69), and *NE* and *IL* (0.66). This indicates that poor reversibility, low necessity, and high information loss tend to co-occur, amplifying each other's effect on the prediction of problematic interventions. *EI* shows negligible interactions with other dimensions (*correlations* $\leq$ 0.22).

To illustrate individual-level explanations, Figure 11 presents two contrasting cases. In the low-risk sample, *RE* (3), *NE* (2), and *IL* (3) all contribute negative SHAP values, pushing

the prediction toward zero. In the high-risk sample, *RE* (5), *NE* (5), and *IL* (5) produce large positive SHAP values, driving the prediction to 0.685. *EI* contributes negligibly in both cases.

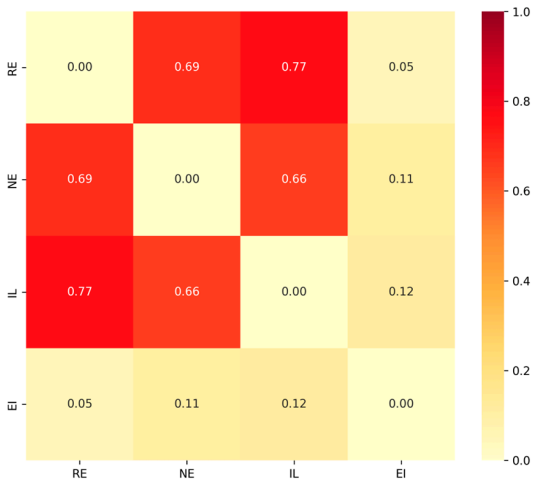
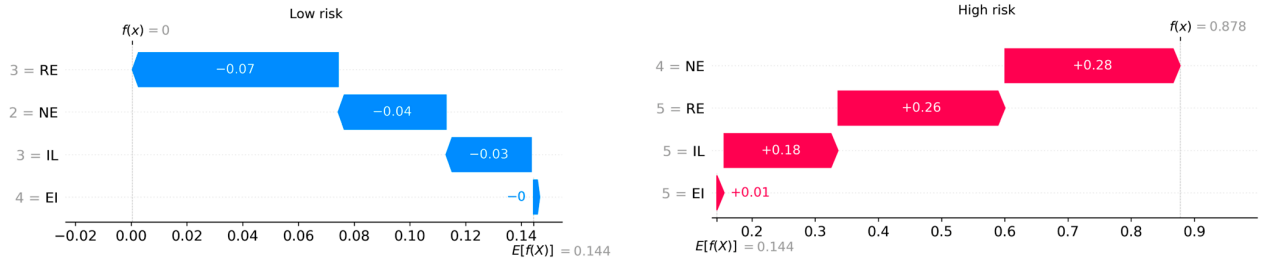


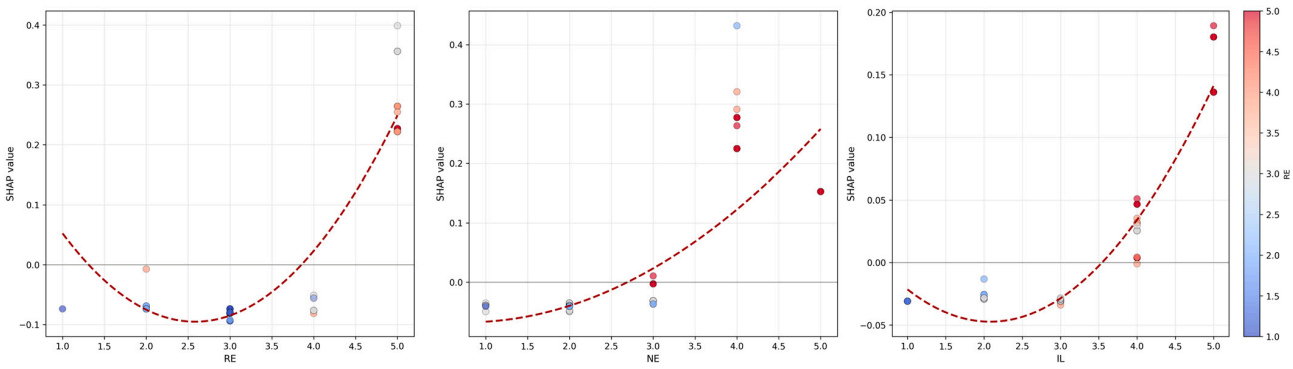
Figure 10. SHAP interaction strength heatmap.



(**a**) (**b**)

Figure 11. Waterfall plots for representative interventions: (**a**) low-risk intervention; (**b**) high-risk intervention.

Finally, Figure 12 visualises the marginal effect of each feature and its interaction with others. *RE* shows a clear threshold: *scores* ≤ 3 yield negative SHAP, while *scores* > 3 cause rapidly increasing positive SHAP, with strong interaction with *NE* ($R^2 = 0.934$). *NE* exhibits a near-linear positive relationship, and *IL* shows a steep rise above score 4, interacting strongly with *RE* ($R^2 = 0.940$). These plots quantify the specific risk intervals for each dimension, providing actionable guidance for conservation planning.



(**a**) (**b**) (**c**)

Figure 12. Advanced dependence plots of SHAP values for the top three features: (**a**) Reversibility (RE), colored by Necessity (NE); (**b**) Necessity (NE), colored by Reversibility (RE); and (**c**) Information Loss (IL), colored by Reversibility (RE).

4. Discussion

This study developed a quantitative framework for minimum intervention assessment by integrating the entropy weight method, random forest classification, and SHAP interpretation, and applied it to 133 documented interventions at Caishen Temple in Xinzhou. The results show a divergence between the entropy-based weight ranking and the random forest-based discriminative importance ranking, provide spatiotemporal evidence of intervention-intensity patterns, and characterise the dimensions most strongly associated with problematic interventions in this dataset. The following sections interpret these findings, compare them with the existing literature, discuss methodological contributions and limitations, and outline implications for conservation practice and future research.

4.1. Divergence Between Data Dispersion and Discriminative Capacity

One of the most striking findings is the substantial difference between the entropy-based weight ranking ($IL > EI > NE > RE$) and the random forest feature importance ranking ($RE > NE > IL > EI$), with a Spearman's ρ of -0.80 ($p = 0.20$, $n = 4$). This divergence highlights a meaningful methodological distinction rather than an error. Entropy weights reflect the relative dispersion of scores across all 133 interventions, essentially capturing which dimensions exhibit the greatest variation among conservation actions. Information Loss (*IL*) and Extent of Intervention (*EI*) showed the highest dispersion, meaning that interventions varied most widely in how much original material was lost or replaced. In contrast, random forest importance identifies which dimensions are most effective at distinguishing the 19 problematic interventions (labeled as inappropriate alterations) from the 114 normal ones. Here, Reversibility (*RE*) and Necessity (*NE*) emerged as the strongest predictors. However, because the comparison involves only four dimensions (n = 4), this rank correlation is presented descriptively and no inferential claim about statistical significance is warranted.

This distinction has important implications for heritage conservation assessment. If one were to rely solely on entropy weights (or any data-driven weighting that captures overall variance), the conclusion would be that *IL* and *EI* should be prioritized in intervention evaluation. However, the random forest analysis indicates that RE and NE have the strongest discriminative association with problematic cases in this dataset. This does not establish a causal relationship; rather, it reflects that among the four correlated dimensions, RE and NE best separate the 19 problematic from 114 normal interventions in this single-site sample. In practical terms, an intervention might exhibit moderate information loss (e.g., replacing some deteriorated tiles) and a moderate extent of replacement, yet still be considered appropriate if it is reversible and necessary. Conversely, an intervention that is irreversible and unnecessary will almost always be judged inappropriate, even if its extent is small and information loss limited. This aligns with the core tenet of minimum intervention: necessity and reversibility are ethical and strategic constraints, whereas extent and information loss are more context-dependent outcomes.

The divergence also highlights a general caution for heritage researchers applying multi-criteria decision-making methods (AHP, entropy, TOPSIS, etc.). These methods are valuable for capturing overall variability or expert preferences, but they do not automatically identify which dimensions are most predictive of undesirable outcomes. When labeled data on "good" vs. "bad" interventions are available (as in this study), supervised learning can provide an exploratory comparison between the derived weights and the dimensions' discriminative importance.

4.2. Spatiotemporal Patterns and Their Conservation Implications

The intervention intensity index (*III*) revealed clear temporal and spatial patterns that resonate with documented conservation histories. The 1992 interventions—carried out during a period when the temple was repurposed by the Health Bureau—exhibited the highest *III* values, with a median near 5 and 100% of measures classified as high-intensity. These interventions were characterised by low necessity (functional conversion rather than structural need) and extremely poor reversibility (cement hardening, layout changes, surface covering). This pattern reflects what might be termed institutional over-intervention: when a heritage asset is assigned to a non-specialist institution without a binding conservation mandate, functional adaptation often overrides preservation ethics. Comparable cases have been documented in other contexts, such as the conversion of historic churches into warehouses or community halls without reversible interventions [1,2]. The 1992 experience at Caishen Temple serves as a cautionary example of how changes in ownership or use can trigger irreversible damage, reinforcing the need for legally enforceable conservation management plans.

In contrast, the 2017 interventions showed markedly lower *III* values (median 2.8) and a broader distribution across core structural components. This shift likely reflects improved conservation awareness, professional oversight, and adherence to the minimum intervention principle. However, the heatmap (Figure 4) shows that even in 2017, some high-intensity measures were applied to beams and columns, suggesting that structural safety concerns sometimes justify larger interventions. The presence of a mixed pattern (low, medium, high intensities on the same component category) indicates that conservators made case-by-case decisions rather than applying a uniform rule: a sign of professional maturity.

The component-level analysis identified roof (36.8% high, 7/19) and walls (33.3% high, 7/21) as the categories with the highest proportion of high-intensity interventions, while decorative finishes (mean 2.11, 0% high) did not exceed the $III \geq 3.5$ threshold. However, beam-frame (mean 2.98, 31.2% high) showed a high-intensity proportion comparable to that of walls (33.3%), suggesting that structural interventions were not inherently lower in intensity. The anticipated dichotomy between 'structural' and 'non-structural' components was not strongly supported by the data, as both categories spanned a similar range of III values. Future training should address both structural and non-structural components, as over-intervention risk appears to depend more on case-specific factors than on the structural role of the element.

4.3. Methodological Contributions and Comparison with Previous Work

The proposed Intervention Intensity Index (*III*) differs from existing quantitative assessment tools in heritage conservation. Most previous indices focus on material decay or structural vulnerability, such as the GNDT form for masonry buildings or the SVIVA method for vulnerability assessment [11,24,25]. These tools evaluate how damaged a building is. The *III*, by contrast, evaluates how invasive an intervention is. It addresses a different but equally important question: given a set of conservation actions, which ones comply with the minimum intervention principle? This shift from damage assessment to intervention assessment is, to the best of our knowledge, novel in the heritage conservation literature.

The combination of entropy weighting and random forest analysis provides a replicable framework for comparing two distinct forms of information: dispersion-based weights derived from the expert-score matrix and discriminative importance estimated against the problematic-intervention labels. In this study, the random forest serves as an exploratory cross-check rather than an independent validation of the entropy weights. The value of

the combined approach therefore lies in examining whether the dimensions that vary most across interventions are also those that best distinguish problematic from normal cases. SHAP further improves model interpretability by showing how the four dimensions contribute to predictions and interact within this dataset.

Compared with recent deep learning-based pathology detection studies, our work takes a different but complementary direction [4,6,8]. Those studies excel at automating the identification of deterioration patterns (cracks, efflorescence, water stains) on building surfaces. Our framework does not detect deterioration; it evaluates the appropriateness of the human response to deterioration. In practice, the two approaches can be integrated: AI-based detection systems can first map decay distributions, and the *III* can then be applied to proposed or executed repair measures to screen for over-intervention. This integration is a promising avenue for future decision-support systems.

4.4. Practical Implications for Conservation Decision-Making

The *III* and the SHAP-derived results may provide preliminary support for conservation planning. The finding that RE and NE showed the strongest discriminative associations with problematic cases in this dataset suggests that heritage managers may benefit from paying particular attention to these two dimensions during the design phase. This interpretation is preliminary and should be tested on additional heritage sites before being adopted as a general rule. For any proposed intervention, two questions should be asked: (1) Is it strictly necessary, or could it be deferred or reduced? (2) Is it reversible, or will it permanently alter the historic fabric? If the answer to the first is "no" and to the second is "irreversible", the intervention should be rejected or redesigned, regardless of its anticipated extent or information loss. This is a more actionable rule than simply aiming for "low *III*".

Second, the cumulative contribution curve (Figure 9) shows that three dimensions (*RE*, *NE*, *IL*) account for over 90% of predictive contribution, while *EI* contributes negligibly. This suggests that assessment tools could be streamlined to focus on the most discriminative dimensions, potentially reducing data collection burden. However, omitting EI is not recommended without external validation, as EI is conceptually central to the minimum intervention principle and its contribution may differ in other datasets.

Third, the threshold effects identified in Figure 12 provide clear numerical anchors for evaluation. A *RE* score of 3 (conditionally reversible, e.g., lime mortar) appears to be a tipping point: below 3, SHAP is negative (protective), and above 3, it becomes increasingly positive (risk-increasing). For *NE*, the relationship is near-linear, so any deviation from "absolutely necessary" (score 1) increases risk. For *IL*, scores above 4 cause sharp risk increases. These thresholds can be incorporated into checklists or semi-automated decision tools.

4.5. Limitations and Future Research

Several limitations should be acknowledged. First, the dataset is limited to a single heritage site (Caishen Temple) with 133 interventions. Although this is a relatively large and well-documented sample for a single building, the generalisability of the weight rankings to other sites remains untested. Different heritage typologies, construction materials, and conservation traditions might yield different importance orders. Future work should apply the *III* framework to multiple sites, including brick, stone, and timber structures, and examine whether the dominance of RE and NE holds across contexts.

A second set of limitations concerns the machine learning validation. The random forest analysis in this study is not an independent validation but an exploratory cross-check, as both the entropy weights and the random forest model were derived from the same

expert-scoring matrix. Moreover, the target variable shares statistical dependence with the input dimensions, as confirmed by Table 5 (all $p < 0.001$, Cohen d > 1.4). To assess stability, repeated five-fold stratified CV (50 repeats) was performed, yielding mean AUC = 0.9485 (95% CI: [0.9361, 0.9595]). The four input dimensions are strongly intercorrelated (EI-IL: 0.91, RE-IL: 0.88, RE-NE: 0.73), which inflates impurity-based feature importance for correlated features [21,22]. Permutation importance, which is more robust to collinearity [21,22], confirms the ranking ($RE > NE > IL > EI$) but with reduced magnitudes, suggesting the ranking should be interpreted with caution. The target variable was defined through a dual-source labelling procedure combining image-based similarity assessment for interventions with paired *pre*- and *post*-restoration photographs and explicit textual evidence from conservation records for interventions without paired photographs. Although this procedure reduced direct dependence on the expert-scoring framework, incomplete photographic coverage and the interpretive nature of the textual evidence mean that the target variable cannot be regarded as fully independent. Future work should validate these findings on larger, more balanced datasets with independent predictor and outcome sources, and prioritise systematic pre- and post-restoration photographic documentation.

Additional limitations relate to the expert scoring and the index formulation. The expert scoring, while subjected to ICC reliability testing, inevitably carries subjective elements. The definition of 'reversibility' and 'necessity' can vary among conservators, and implicit biases may remain despite the standardised rubric. Future studies could incorporate larger and more diverse expert panels or formalise RE and NE using engineering criteria. The *III* calculation currently relies on expert scores, which could be automated in the future using natural language processing of conservation reports. Furthermore, the study does not incorporate post-intervention monitoring data, and the SHAP analysis revealed strong interactions among RE, NE, and IL, suggesting that a non-linear composite index could be explored in future work.

5. Conclusions

This study developed a quantitative framework for minimum intervention assessment in architectural heritage conservation, integrating the entropy weight method, random forest classification, and SHAP interpretation. Applied to 133 documented interventions at Caishen Temple in Xinzhou across four periods (1992, 2006, 2016, 2017), the framework demonstrated that the minimum intervention principle can be operationalised as a measurable Intervention Intensity Index (*III*). The entropy weight method assigned the highest weights to Information Loss (0.3149) and Extent of Intervention (0.3025), reflecting their overall dispersion across all interventions. However, the random forest classifier ($AUC = 0.9485$, 95% CI: [0.9361, 0.9595]) identified Reversibility (mean impurity importance = 0.403) and Necessity (mean impurity importance = 0.370) as the most discriminative features—a divergence (Spearman's $\rho = -0.80$, $n = 4$, descriptive only) that suggests that what varies most among interventions is not necessarily what most strongly differentiates appropriate from inappropriate practice. SHAP analysis further revealed threshold effects ($RE\ score > 3$, $IL\ score > 4$) and strong synergistic interactions among *RE*, *NE*, and *IL*, confirming that poor reversibility, low necessity, and high information loss tend to co-occur and amplify the risk of over-intervention. Spatiotemporally, interventions from the 1992 non-specialist occupancy period exhibited the highest *III* values ($median \approx 5$), concentrated on non-structural walls and floors, while the 2017 interventions showed markedly lower intensity ($median \approx 2.8$) and broader coverage of core structural components, reflecting improved conservation practice. Methodologically, the *III* shifts the focus from assessing material decay to evaluating the invasiveness of human responses to decay—a novel contribution that complements existing pathology detection tools. Practically, the framework

provides a case-based prototype that demonstrates how the *III* framework can be applied to compare repair alternatives and screen proposed interventions. Further validation on additional heritage sites is needed before generalising the framework as a routine tool. Limitations include the single-site dataset and the inherent subjectivity of expert scoring. Future work should test the framework across multiple heritage typologies, automate dimension scoring using natural language processing, and incorporate long-term post-intervention monitoring to calibrate *III* thresholds against actual durability outcomes. With these extensions, the proposed framework may evolve from a research prototype into a decision-support tool for sustainable heritage conservation, subject to further validation across diverse heritage contexts.

Author Contributions: Conceptualization, T.L. and S.S.B.H.; methodology, T.L. and S.S.B.H.; validation, G.Z., S.S.B.H. and Y.W.; formal analysis, T.L.; investigation, T.L., G.Z. and Y.W.; resources, G.Z. and Y.W.; data curation, T.L. and G.Z.; writing—original draft preparation, T.L.; writing—review and editing, T.L., G.Z., S.S.B.H. and Y.W.; visualization, T.L.; supervision, S.S.B.H.; project administration, T.L.; funding acquisition, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 2026 industry–university collaborative project between Jiaozuo University and Tianjin Xinfei Digital Technology Co., Ltd., grant number JDHX2026-001. The APC was funded by Jiaozuo University.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee for Research Involving Human Subjects, University Putra Malaysia (protocol code JKEUPM-2024-837 and 7 October 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The complete code and anonymised dataset are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to express their sincere gratitude to the conservation authorities and site managers of Caishen Temple in Xinzhou for providing access to the conservation archives and facilitating the on-site survey. The authors also thank the expert scorers, whose professional backgrounds cover historic building restoration, architectural heritage conservation, materials science, and field documentation, for their careful and independent evaluation of the documented intervention measures.

Conflicts of Interest: Author Yang Wang was employed by the company Tianjin Xinfei Digital Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Tianjin Xinfei Digital Technology Co., Ltd. The funder had the following involvement with the study through its employee and co-author Yang Wang: validation, investigation, provision of resources, and writing—review and editing.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Receiver Operating Characteristic Curve
EI	Extent of Intervention
ICC	Intraclass Correlation Coefficient
IL	Information Loss
NE	Necessity
NLP	Natural Language Processing
RE	Reversibility

RF	Random Forest
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SHAP	Shapley Additive Explanations
GNDT	Gruppo Nazionale per la Difesa dai Terremoti

References

1. Feilden, B.M. *Conservation of Historic Buildings*, 3rd ed.; Architectural Press: Oxford, UK, 2003. [CrossRef]
2. ICOMOS. *Principles for the Conservation of Wooden Built Heritage*; ICOMOS: Charenton-le-Pont, France, 2017. Available online: https://www.icomos.org/images/DOCUMENTS/General_Assemblies/19th_Delhi_2017/Working_Documents-First_Batch-August_2017/GA2017_6-3-4_WoodPrinciples_EN_final20170730.pdf (accessed on 24 May 2026).
3. Zhang, Y.; Dong, W. Determining minimum intervention in the preservation of heritage buildings. *Int. J. Archit. Herit.* **2021**, *15*, 698–712. [CrossRef]
4. Yan, L.; Chen, Y.; Zheng, L.; Zhang, Y. Application of computer vision technology in surface damage detection and analysis of shedthin tiles in China: A case study of the classical gardens of Suzhou. *Herit. Sci.* **2024**, *12*, 72. [CrossRef]
5. Chen, X.; Wang, J.; Wang, S.; He, Y.; Cao, M. Intelligent Detection and Virtual Restoration of Shedthin Tiles in Suzhou's Classical Gardens. *Int. J. Archit. Herit.* **2025**, 1–21. [CrossRef]
6. Zhang, Y.; Kong, L.; Antwi-Afari, M.F.; Zhang, Q. An Integrated Method Using a Convolutional Autoencoder, Thresholding Techniques, and a Residual Network for Anomaly Detection on Heritage Roof Surfaces. *Buildings* **2024**, *14*, 2828. [CrossRef]
7. Zou, J.; Deng, Y. Intelligent Assessment System of Material Deterioration in Masonry Tower Based on Improved Image Segmentation Model. *Herit. Sci.* **2024**, *12*, 252. [CrossRef]
8. Gbran, H. A Hybrid Machine-Learning Framework for 3D Chromatic Decay Detection and Spatial Mapping in Tropical Heritage Buildings. *Int. J. Archit. Herit.* **2026**, 1–29. [CrossRef]
9. Al-Omari, A.; Khattab, S. Characterization of Building Materials Used in the Construction of Historical Al-Omariya Mosque Minaret in Mosul's Old City, Iraq. *J. Build. Eng.* **2021**, *33*, 101645. [CrossRef]
10. Wu, R.; Hao, Y.; Xuan, J.; Hou, Z.; Feng, W. Multi-Factorial Damage Mechanisms of Chinese Grey Bricks under Complex Natural Environments: A Comprehensive Review from Manufacturing to Performance Degradation. *Int. J. Archit. Herit.* **2026**, 1–33. [CrossRef]
11. Sánchez-Calvillo, A.; Ortega Heras, J.; Alonso Guzmán, E.M. Seismic Vulnerability Assessment of Mexican Vernacular Adobe Buildings: Application and Comparison of Different Methods. *Int. J. Archit. Herit.* **2025**, *19*, 2897–2918. [CrossRef]
12. Fiore, P.; Sicignano, E.; Donnarumma, G. An AHP-Based Methodology for the Evaluation and Choice of Integrated Interventions on Historic Buildings. *Sustainability* **2020**, *12*, 5795. [CrossRef]
13. Maselli, G.; Cucco, P.; Nesticò, A.; Ribera, F. Historical Heritage–Multi-Criteria Decision Method (H-MCDM) to Prioritize Intervention Strategies for the Adaptive Reuse of Valuable Architectural Assets. *MethodsX* **2024**, *12*, 102487. [CrossRef] [PubMed]
14. Zhang, Y. Minimum Intervention Level Decision for Historical Buildings: Historical Buildings along the Central Axis of Beijing, China. *npj Herit. Sci.* **2026**, *14*, 50. [CrossRef]
15. Liu, F.; Lu, Z.; Qiang, W. An Innovative Approach to Post-Occupancy Evaluation of Architectural Heritages in the Peking-Mukden Railway. *Herit. Sci.* **2024**, *12*, 383. [CrossRef]
16. Bertolin, C.; Loli, A. Sustainable Interventions in Historic Buildings: A Developing Decision Making Tool. *J. Cult. Herit.* **2018**, *34*, 291–302. [CrossRef]
17. Cardinale, T.; Colapietro, D.; Cardinale, N.; Fatiguso, F. Evaluation of the Efficacy of Traditional Recovery Interventions in Historical Buildings: A New Selection Methodology. *Energy Procedia* **2013**, *40*, 515–524. [CrossRef]
18. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
19. Shanxi Ancient Architectural Design Co., Ltd. Conservation Design Scheme for Caishen Temple in Xinfu District, Xinzhou, Shanxi, China. 2017; *unpublished internal document*.
20. Likert, R. A Technique for the Measurement of Attitudes. *Arch. Psychol.* **1932**, *22*, 5–55. Available online: https://legacy.voteview.com/pdf/Likert_1932.pdf (accessed on 5 July 2026).
21. Kelleher, J.D.; Mac Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2020. Available online: https://mitpress.mit.edu/9780262044691/fundamentals-of-machine-learning-for-predictive-data-analytics/ (accessed on 5 July 2026).
22. Harrou, F.; Zeroual, A.; Hittawe, M.M.; Sun, Y. Traffic congestion detection: Data-based techniques. In *Road Traffic Modeling and Management: Using Statistical Monitoring and Deep Learning*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 141–195. [CrossRef]

23. Gong, C.; Zeng, S.; Liu, D. Quantitative Analysis of Comprehensive Similarity in Restoration of Ancient Building Walls Using Hue–Saturation–Value Color Space and Circular Local Binary Pattern. *Buildings* **2024**, *14*, 1478. [CrossRef]
24. Ceroni, F.; Caterino, N.; Vuoto, A. Simplified seismic vulnerability assessment methods: A comparative analysis with reference to regional school building stock in Italy. *Appl. Sci.* **2020**, *10*, 6771. [CrossRef]
25. Ortega, J.; Vasconcelos, G.; Rodrigues, H.; Correia, M. A vulnerability index formulation for the seismic vulnerability assessment of vernacular architecture. *Eng. Struct.* **2019**, *197*, 109381. [CrossRef]

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.