Examining the Relationships Between Historic Building Features and Tornado Damage: A Multi-Model Feature Importance Analysis with Statistical Validation

Saanchi S. Kaushal[a], Yishuang Wang[a], Mariantonieta Gutierrez Soto, M.ASCE[b], Rebecca Napolitano, M.ASCE[a,1]

[a]*Dept. of Architectural Engineering, Pennsylvania State University, University Park, PA 16802, United States*

[b]*School of Engineering Design and Innovation, The Pennsylvania State University, 307 Engineering Design and Innovation Bldg., University Park, PA 16802, United States*

Abstract

Historic masonry buildings constitute significant cultural and economic assets including tornado-prone regions, yet their vulnerability characteristics remain poorly quantified. This study analyzes damage patterns in 382 historic structures exposed to the 2020 Nashville (EF3/EF4) and 2021 Quad State (EF4) tornado outbreak, employing permutation importance and SHapley Additive exPlanations (SHAP) analysis to identify building features governing tornado vulnerability. Feature importance rankings were validated against random noise controls to establish statistical significance. Two modeling approaches reveal distinct insights: hazard-inclusive models (including EF rating) demonstrate that tornado intensity overwhelmingly dominates damage prediction, suppressing building feature effects; hazard-neutral models (excluding intensity) elevate intrinsic structural characteristics, with wall thickness, roof slope, and construction year emerging as top predictors.

Rather than providing prescriptive design specifications, this work establishes a evaluation framework identifying high-priority targets for wind tunnel testing, component-level experimentation, and finite element modeling. The analysis emphasizes preservation-compatible interventions that balance life safety with architectural integrity, guiding resource allocation for the vulnerable historic building stock.

Keywords: Tornado Damage, Feature Importance, Historic Buildings, Machine Learning, Statistical Validation, Permutation Importance

*Corresponding author. Email: nap@psu.edu

1. Introduction

The preservation of historic building stock faces an existential threat due to the increasing frequency and intensity of severe convective storms. There are over 95,000 historic buildings in the United States [1], they often act as the economic and cultural anchors of their communities, are particularly vulnerable to tornado-induced wind loads due to construction practices that predate modern engineering codes [2]. The devastation of Mayfield, Kentucky's historic district during the 2021 Quad State tornado outbreak serves as a stark reminder of this fragility [3]. While modern building codes have evolved to improve life safety, historic structures, often characterized by unreinforced masonry (URM) and gravity-based connections [4], occupy a precarious position where standard engineering interventions may conflict with preservation mandates for material integrity and reversibility [5].

A challenge in mitigating this risk is the lack of empirical data linking specific historic building features to tornado performance. Preservation professionals often rely on anecdotal evidence or generalized wind engineering principles that may not fully capture the complex failure mechanisms of aged structures. Furthermore, the "transition zone" of damage, where buildings sustain repairable structural damage without progressing to total loss, remains poorly understood, yet this is precisely the domain where preservation interventions are most valuable. This intermediate damage state, characterized by partial roof loss, wall cracking, or localized collapse that can be addressed through structural restoration rather than demolition, represents the critical threshold determining whether historic structures can be saved following tornado impact. Unlike undamaged buildings (requiring no intervention) or completely destroyed structures (beyond repair), transition zone buildings face uncertain futures where preservation decisions depend on accurate damage assessment and targeted retrofitting strategies.

To address this knowledge gap, the study implements an exploratory machine learning analysis of post-tornado damage data. The primary objective is not to develop a predictive black-box model for automated assessment, but rather to use interpretable machine learning techniques to generate testable hypotheses regarding historic building vulnerability. The key question being which observable features differentiate structures that survive from those that experience significant damage in historic-style construction. Clarifying these differences supports a more targeted and impactful use of limited preservation resources for engineering assessments and retrofit planning.

The authors also recognize the limitations of the dataset, particularly the small sample size of "low damage" cases (n=20) and the inherent circularity of damage-

based EF ratings. To tackle this, the research utilizes statistical equivalence testing and a random noise guardrail to filter out spurious correlations. Permutation Importance was implemented for global feature ranking [6] and integrated with SHAP (SHapley Additive exPlanations) analysis [7], computed on held-out validation data to explore potential interaction effects. This approach moves the study beyond simple correlation to identify mechanistic candidates for future investigation, such as the compounding risk of specific wall-roof combinations. In doing so, the study established a data-driven foundation for more nuanced discussions about risk, resilience, and the limits of intervention in the historic built environment. The analytical framework prioritizes methodological safeguards that prevent spurious findings, recognizing that preservation decisions based on unreliable correlations could lead to ineffective or counterproductive interventions.

The proposed approach incorporates several methodological innovations that distinguish it from conventional disaster assessment studies. Instead of reporting results from a single purportedly optimal model, a practice that risks overfitting to dataset idiosyncrasies, the analysis benchmarks six model families using statistical equivalence testing to identify all models whose performance is indistinguishable from the best. This multi-model validation ensures that identified vulnerabilities replicate across different analytical approaches, substantially increasing confidence in the findings. Additionally, a synthetic random feature is introduced as a negative control. If this noise variable appears among the important features, the validity of the analysis is called into question [8]. This provides an objective quality check that is largely absent from existing disaster assessment studies.

The combination of permutation importance analysis, which provides global feature rankings, with SHAP analysis, which reveals instance-level mechanisms and feature interactions, proves particularly valuable for historic preservation applications. While permutation importance identifies which features matter across the entire building stock, SHAP analysis reveals how these features combine in specific buildings, enabling preservation professionals to identify structures facing compounded risk from multiple vulnerabilities. For instance, SHAP analysis reveals that wall substrates and building height both push predictions toward significant damage, though this cannot quantify whether their combined effect is additive or multiplicative without formal interaction testing. Such interaction effects are important for retrofit prioritization, as addressing either vulnerability factor in isolation provides limited benefit compared to holistic interventions.

The results from this study demonstrate that feature importance analysis can yield valid scientific insights even when predictive model performance appears modest by conventional machine learning standards. While the macro F1 score

(0.48) reflects genuine difficulty in predicting the rare transitional “Low" damage class, the models successfully identify buildings at the highest collapse risk (F1=0.72 for Significant Damage), the outcome most relevant for preservation prioritization. This distinction between predictive accuracy and feature importance reliability carries implications for disaster assessment research, where perfect prediction often proves impossible, yet actionable insights remain achievable.

2. Data

2.1. Tornado Events and Data Collection

The analysis draws upon building-level damage data from two major tornado events. The first event occurred on March 3, 2020, when an EF3-EF4 tornado carved a 25-mile track through Nashville, Tennessee [9], while the second event, the December 10-11, 2021 Quad State tornado, produced a long-track EF4 tornado affecting Kentucky, Tennessee, Arkansas, and Missouri [10]. Following the tornadoes, the Structural Extreme Events Reconnaissance Network (StEER) deployed the Virtual Field Assessment Team (VAST) and the Field Assessment Structural Team (FAST) for Nashville [9] and Quad State [10] to document the extent of damages. The authors participated in both the field deployments.

Damage was evaluated for 382 buildings across both tornado events, of which 230 were listed in the *National Register of Historic Places*. Data collection employed two complementary approaches: on-site field reconnaissance for structures with safe access, and virtual assessment using pre- and post-event remote sensing for buildings where physical access was restricted or pre-tornado documentation was available.

Field reconnaissance focused on Mayfield’s downtown historic district [11], which sustained catastrophic damage during the Quad State tornado outbreak. The district’s concentration of buildings constructed between 1850 and 1930 provided a unique opportunity to study tornado vulnerability in historic masonry structures predating wind engineering codes. On-site data collection utilized the Fulcrum application for standardized damage surveys, DJI Matrice drones for aerial documentation, and Street View cameras for façade capture [3, 12].

Virtual reconnaissance extended the dataset beyond field-accessible buildings, evaluating all historic structures within a 2-mile radius of tornado paths for both the Quad State [13] and Nashville [12] events. High-resolution aerial imagery (5–7.5 cm ground sample distance) from Nearmap [14] and Google Earth [15] provided roof-level details including geometry, slope, and covering materials.

Street-level panoramas from Google Street View captured façade conditions, fenestration patterns, and cladding systems. Pre-event imagery proved particularly valuable for documenting original construction details—roof substrates, wall materials, parapet heights, connection types—subsequently destroyed or obscured by damage, enabling retrospective assessment of as-built conditions influencing tornado performance.

Data collection followed standardized StEER protocols [16], systematically documenting structural attributes, geometric properties, and damage indicators for each building. Structural features captured included construction type (masonry unreinforced, wood frame, hybrid systems), wall substrate and thickness, foundation type, and Main Wind Force-Resisting System (MWFRS) configuration. Roof characteristics received particular emphasis given their documented influence on tornado performance [17], with assessors recording roof shape (gable, hip, flat, mansard), slope, substrate material, covering type, and overhang dimensions. Wall and fenestration details documented cladding materials, opening percentages on all elevations, and parapet heights where present, as these features influence both structural capacity and internal pressurization following envelope breach [18].

Field and remote observations were supplemented with archival data from the National Register of Historic Places (NRHP) database, providing construction dates, architectural styles, and documentation of previous alterations or retrofits. The combined dataset comprised 382 buildings constructed between 1850 and 1950 across both tornado events. Temporal distribution of the building stock (Figure 1) confirms the historic character, with peak construction years between 1890 and 1930. The Quad State sample exhibits a longer tail of pre-1880 structures compared to Nashville, reflecting Mayfield's older urban core. Of the 382 buildings, 230 held formal historic designation through the National Register of Historic Places or local historic districts (10 Nashville, 220 Quad State), while the remaining 150 buildings exhibited similar construction characteristics (masonry bearing walls, timber roof framing, shallow foundations), but lacked formal designation, representing the vernacular historic building stock vulnerable to tornado damage.

2.2. Dataset Characteristics

The combined dataset comprises 382 historic masonry buildings spanning construction years 1850–1950, exposed to EF-scale tornado intensities ranging from EF0 to EF4. While all buildings share fundamental characteristic of masonry construction, substantial variation exists across structural, geometric, material, and hazard dimensions that govern tornado vulnerability.

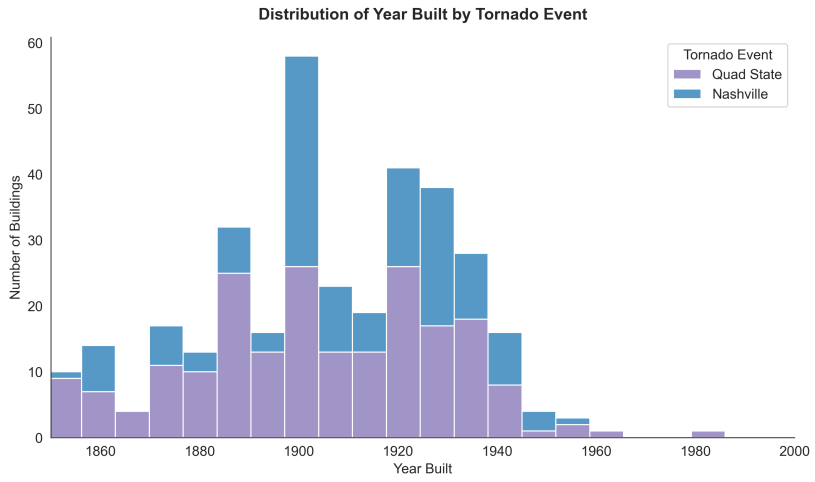


Fig. 1. Distribution of building construction years by tornado event. The dataset is concentrated between 1890 and 1930, with Quad State including a tail of older (pre-1880) structures.

Structural attributes include number of stories (1–4), roof shape (gable, hip, flat, mansard), foundation type (rubble stone, brick pier, continuous wall), and MWFRS configuration distinguishing frame-based versus masonry diaphragm lateral systems. Geometric properties vary across building footprint area (40–5,630 m^2), height (3.5–23.3 m), wall dimensions (side length 4.9–92 m, front length 5.5–211 m), and roof slope (0–60°). Material characteristics capture heterogeneity through wall thickness (0.2–0.6 m), wall substrate type (masonry unreinforced, reinforced, wood frame), roof substrate (dimensional lumber, trusses, wood sheathing grade), roof covering (asphalt, slate, clay tile, metal), and cladding systems (brick veneer, wood siding, stucco). Envelope features document fenestration percentage on each elevation (0–90%), parapet height (0–1.5 m), and overhang length (0–5 m), all influencing wind pressure distribution and internal pressurization risk.

Hazard variables quantify tornado exposure through EF rating assigned to each building location and distance from tornado path centerline (0–2 km). The distribution of EF ratings (Figure 2) reflects differing intensities between events: the Quad State tornado contributed higher proportions of EF3 and EF4 exposures, while Nashville exhibited more EF0–EF2 exposures, providing variance necessary to study building performance across the full intensity spectrum. Contextual factors include urban setting (isolated, row-middle, row-end), building position relative to street, occupancy type (residential, commercial, religious, institutional), and historic designation status (NRHP-listed versus visual assessment). This multidi-

mensional feature space enables statistical models to identify vulnerability patterns specific to historic masonry construction while accounting for confounding factors such as building size, occupancy-driven design differences, and spatial clustering within historic districts.

This multidimensional feature space enables statistical models to identify vulnerability patterns specific to historic masonry construction while accounting for confounding factors such as building size, occupancy-driven design differences, and spatial clustering within historic districts.

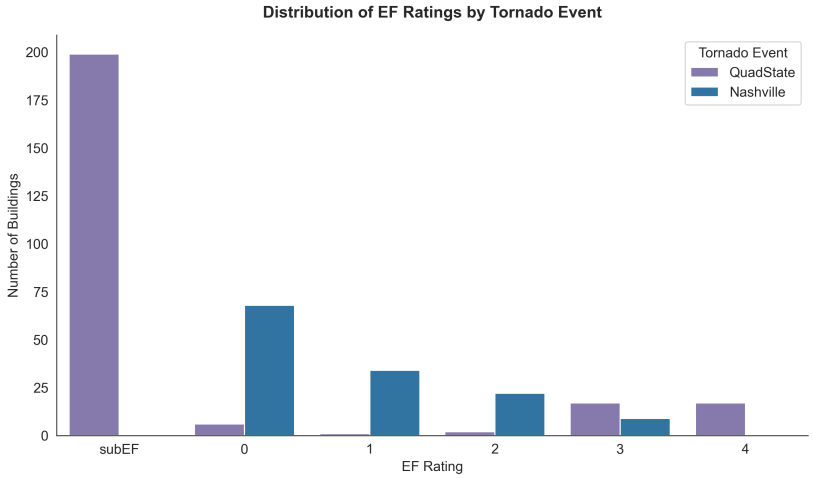


Fig. 2. Distribution of EF ratings in the dataset. The Quad State tornado outbreak event contributes the majority of high-intensity (EF3-EF4) exposures.

2.3. Target Variable and Class Distribution

Damage observed during reconnaissance was classified using the five-category StEER definitions (undamaged, minor, moderate, major, destroyed). However, intermediate damage states contained fewer than 10 observations each, making statistical analysis with this fine-grained classification impractical. The original categories were therefore collapsed into three classes defined by preservation outcomes: Class 0 (Undamaged) requires no intervention, Class 1 (Low Damage) requires repair but preserves historic fabric, and Class 2 (Significant Damage) necessitates major reconstruction or represents total loss. This aggregation strategy follows established practices for handling limited sample sizes and reducing class sparsity [19, 20].

The resulting dataset exhibits severe class imbalance characteristic of post-disaster surveys (Figure 3): 294 buildings (77%) sustained no damage requiring intervention, 41 buildings (11%) experienced repairable damage preserving the majority of historic material, and 47 buildings (12%) suffered severe damage or total loss. This three-class structure prioritizes the critical preservation decision boundary distinguishing structures that can be saved from those facing demolition, rather than imposing artificial distinctions among undamaged or catastrophically damaged buildings where preservation interventions offer no value.

The class imbalance creates methodological challenges for performance evaluation. Simple accuracy would be misleading, as a naïve classifier predicting "undamaged" for all buildings achieves 77% accuracy while providing no useful information. Macro-averaged F1 score is therefore employed as the primary performance metric, as it equally weights performance across all damage classes regardless of sample size, penalizing models that ignore minority classes [21, 22]. This ensures models demonstrate meaningful discriminative capacity for the preservation-critical Low and Significant Damage categories, not merely high accuracy from correctly classifying the dominant Undamaged class.

Analysis of building age versus damage severity (Figure 4) reveals no strong correlation, indicating that construction era alone does not predict vulnerability. Rather, specific structural details (wall thickness, roof-wall connections, retrofit presence) and maintenance conditions likely govern tornado performance independent of building age, motivating the feature-based vulnerability analysis presented in subsequent sections.

Figure 4 shows the distribution of construction years across damage classes. The three boxplots exhibit substantial overlap, with all classes centered around 1900 (median years: Undamaged 1900, Low Damage 1905, Significant Damage 1910). This overlap indicates that building age alone does not reliably predict tornado vulnerability. Older buildings are not systematically more vulnerable than newer ones in this dataset, nor are the newest buildings systematically safer.

This absence of a clear age-damage relationship contradicts the intuitive expectation that older buildings, having experienced more degradation, should perform worse. However, the pattern reflects survivorship bias rather than age-irrelevance. Pre-1880 structures present in the dataset represent a pre-selected cohort of high-quality survivors: poorly constructed contemporaries were demolished decades ago, leaving only robust examples. In contrast, buildings from the 1900–1920 construction age includes development that has not yet been culled by time or economic obsolescence. Additionally, pre-1880 buildings are more likely to hold formal historic designation, receiving preservation-quality maintenance that off-

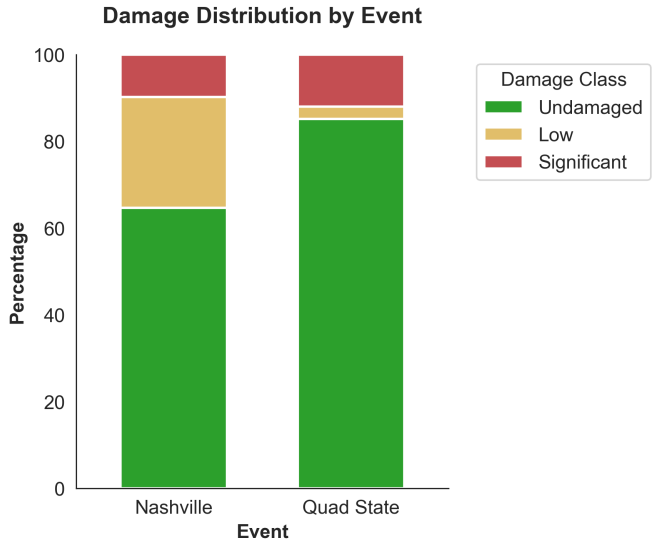


Fig. 3. Proportion of damage classes by tornado event. Quad State shows a higher rate of significant damage due to the direct impact on the historic district.

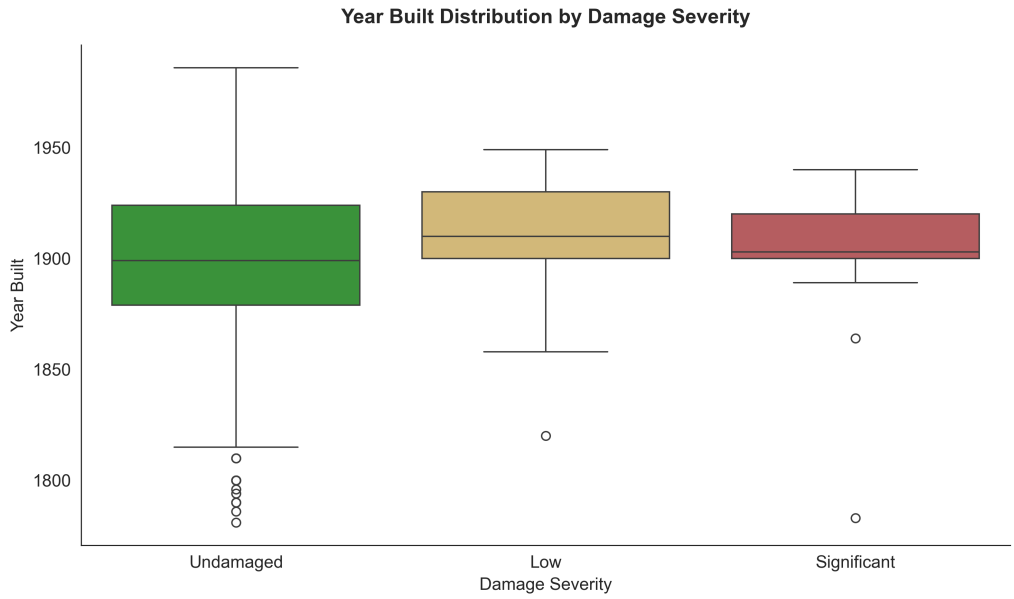


Fig. 4. Boxplot of Year Built by Damage Class. The lack of a clear linear trend suggests age alone is not a strong predictor of vulnerability.

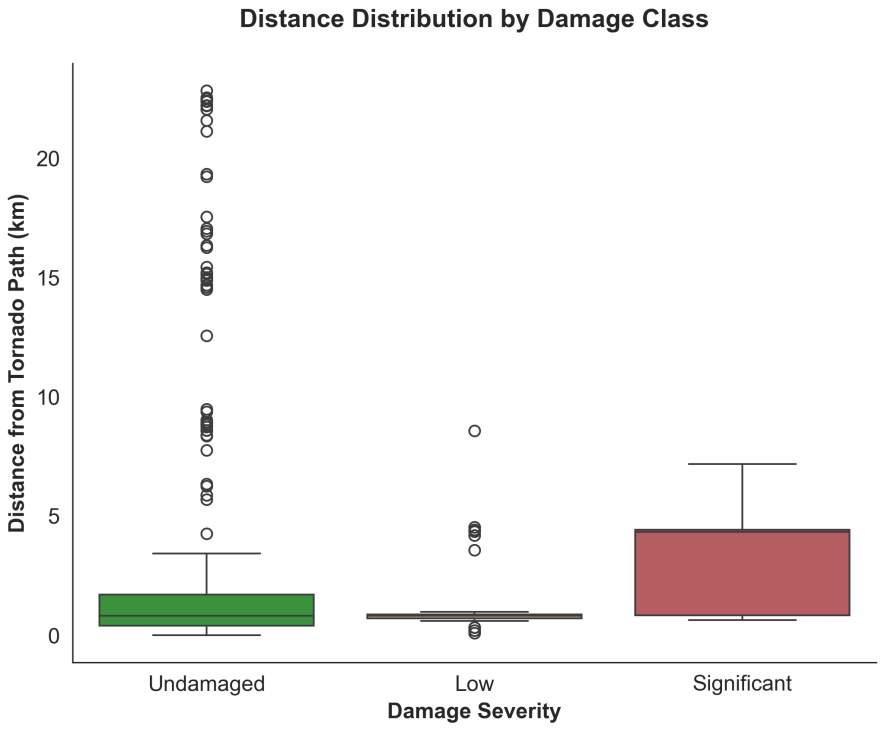


Fig. 5. Boxplot of Tornado Distance by Damage Class. The overlapping distributions suggest distance alone is not a strong predictor of vulnerability.

sets age-related degradation, while early 20th-century buildings may lack both designation protections and consistent owner investment.

The overlapping distributions demonstrate that chronological age functions as a poor vulnerability proxy without accounting for maintenance history, original construction quality, and most critically, tornado exposure variables that govern actual wind loading. Two buildings constructed in 1900 may exhibit drastically different performance if one experienced EF0 winds at 1 km from the tornado path while the other faced EF3 winds at 100 m distance. Age-based screening would misclassify both. This finding motivates the feature-based vulnerability analysis in subsequent sections, which identifies specific structural characteristics (wall thickness, roof geometry, retrofit presence) governing performance independent of construction era.

Additionally, figure 5 reveals the relationship between distance from tornado centerline and damage outcomes. Low damage buildings exhibit the tightest dis-

tribution, defining a narrow "transition zone" where wind speeds cause repairable structural damage without progression to total loss. Significantly damaged buildings show greater median distance and wider dispersion than low damage cases, reflecting the confounding influence of local EF rating variation along the tornado path—buildings 2–3 km from an EF4 centerline can experience EF2–EF3 winds exceeding design thresholds, while buildings 1 km from an EF1 path may sustain only minor damage. Undamaged buildings span the full distance range, with substantial overlap across all damage classes. This overlap demonstrates that distance alone, without accounting for local wind speed and building-specific structural characteristics, provides insufficient damage prediction. The finding motivates hazard-neutral modeling approaches isolating intrinsic building vulnerabilities independent of tornado exposure intensity.

2.4. Wind Vulnerability in Masonry Structures

Unreinforced masonry buildings exhibit tornado vulnerability mechanisms stemming from their fundamental structural characteristics. Historic masonry construction employed load-bearing walls as the primary structural system [4], with two-leaf or three-leaf configurations providing compressive capacity but negligible tensile strength due to masonry's anisotropic material properties [23]. Traditional gravity-based connections between roof and wall elements, while adequate for transferring shear forces under normal loading, prove inadequate for resisting tensile uplift forces generated by tornado wind fields [24].

Failure initiates through several mechanisms that frequently interact to produce progressive collapse. Out-of-plane wall failure occurs when wind pressure causes masonry walls to act as one-way slabs spanning between floor and roof diaphragms; insufficient anchorage to these horizontal elements leads to mid-height cracking and potential wall collapse [25]. Roof uplift results from negative pressure coefficients on leeward and roof surfaces [26], with failure of roof-to-wall connections allowing entire roof system separation. Parapet overturning generates large moments at the base of unbraced masonry elements projecting above the roofline, particularly vulnerable given their exposure to peak wind velocities and lack of lateral restraint. Internal pressurization from breached envelope openings (windows, doors, cladding) can amplify net roof uplift forces [27, 28], transforming moderate external suction into failure-inducing combined loading.

These mechanisms exhibit cascading interdependence. Cladding loss or fenestration breach precedes structural damage through internal pressurization initiation [28]. Roof loss removes lateral bracing for tall masonry walls, triggering secondary out-of-plane collapse even after wind speeds decrease. Parapet failure generates

debris impact loading on adjacent roof and wall surfaces. The spatial heterogeneity of wind pressures across building surfaces, varying with geometry and orientation [29], means simultaneous loading combinations differ substantially from design assumptions based on uniform pressure distributions, creating unexpected stress states that exploit connection and material vulnerabilities inherent to unreinforced masonry construction.

3. Methodology

3.1. Overview of Multi-Model Framework

Following the theoretical framework established by [30], this study prioritizes explanatory modeling over pure prediction, to distinguish between models designed to identify causal mechanisms versus those optimized solely for forecasting. Understanding which building features drive vulnerability is more valuable than achieving marginal gains in aggregate classification accuracy for preservation engineering. Recent empirical work by [31] demonstrates that feature importance rankings remain stable and valid even when model performance degrades, given that the degradation stems from irreducible noise rather than sample size limitations.

For this study, six model families were benchmarked and statistical equivalence testing was implemented to identify all models that perform indistinguishably from the top performer as compared to reporting results from a single optimal model. To ensure that the vulnerabilities identified reflect genuine building characteristics rather than algorithmic artifacts, the findings were replicated across multiple equivalent models. This cross-model validation helped filter out model-specific noise and reduced the risk that preservation recommendations rest on idiosyncratic behavior. As a result, features ranking highly across different algorithmic approaches represent a genuine signal, providing preservation professionals with confidence that retrofit priorities address real vulnerabilities.

A synthetic negative control feature was generated from a standard normal distribution with a fixed random seed and added to the dataset, as established [32] and operationalized in the Boruta algorithm. This random probe contained zero information and provided an objective baseline for statistical significance. Any physical feature (such as parapet height or MWFRS configuration) that consistently outperforms this random baseline across multiple model families is statistically distinguishable from noise, regardless of the model's overall accuracy [32]. This guardrail served two purposes for preservation applications; This random noise feature serves as a quality control mechanism with two functions. First, if random

noise ranks among important predictors, the analysis is flawed—indicating data leakage or spurious correlations that would render any preservation recommendations unreliable. Second, the noise feature provides an objective threshold that any building feature ranking below random noise lacks genuine predictive signal and should be excluded from interpretation, preventing preservation resources from being misdirected toward irrelevant characteristics.

3.2. Data Preparation

Two hazard variables helped quantify tornado exposure: EF rating and distance from tornado path. EF rating represents tornado intensity at each building location that was assigned during damage surveys. These ratings (EF0 through EF5) were encoded as integers 0 through 5, while subEF events (wind speeds below EF0 threshold) were coded as -1. Distance from tornado path represents the shortest distance from each building centroid to the tornado track centerline, calculated as point-to-segment distance using planar approximation with latitude-dependent coordinate scaling [33].

One of the challenges with tornado damage modeling is the potential for circular reasoning when using EF ratings as predictors. Since EF ratings are post-hoc intensity estimates often derived from the building damage [34], including them creates a tautological loop where the outcome (damage) implicitly informs the predictor (EF). To address this, the analysis was conducted under two distinct conditions: First, the Hazard-Neutral approach was considered to represent structural truth. This condition excluded EF rating and distance-to-track, forcing the model to predict damage solely based on intrinsic building characteristics (e.g., geometry, materials, age). This was the primary lens for identifying structural vulnerabilities, as it eliminates the circularity of the EF scale. Second, the Hazard-Inclusive approach was implemented to provide contextual control. This condition includes hazard features to quantify how much predictive power is gained by knowing the wind intensity and distance to the building. While this introduces circularity, it serves as a necessary control to benchmark the relative importance of structural features against the overwhelming force of the winds. The findings are explicitly prioritized from the Hazard-Neutral condition for structural recommendations, treating Hazard-Inclusive results primarily as a validation of the model's ability to capture basic physical reality (i.e., stronger winds cause more damage).

High missingness (>10%) is observed for several features, particularly those requiring interior access or detailed inspection. This missingness is likely informative rather than random: buildings with extensive damage may have had

Table 1. Dataset Composition by Key Features (with Missingness)

Feature	Category	Count (%)	Missing (%)
Number of Stories	1 Story	250 (65%)	0%
	2 Stories	110 (28%)	
	3+ Stories	26 (7%)	
Roof Shape	Gable	280 (73%)	0%
	Hip	60 (16%)	
	Flat	46 (12%)	
Foundation Type	Continuous	300 (78%)	36%
	Pier	50 (13%)	
	Slab	36 (9%)	
Roof Substrate	Board/Plank	304 (80%)	19.6%
Wall Substrate	URM/Brick	182 (48%)	13.1%
Retrofit Type	Present	107 (28%)	8%
Wall Thickness	Mean: 380mm	–	36%
Fenestration (%)	Mean: 15-20%	–	15-20%

inaccessible interiors, while remote-sensing-only assessments could not document hidden details. For retrofit-related features, missing data often indicates "no retrofit was documented during reconnaissance," which can mean there was no retrofit or inaccessible due to building collapse. Preprocessing preserved this information through two strategies, median imputation for numeric categories and ordinal encoding for categorical features. This allows models to learn whether "unknown" status correlates with damage. For instance, if buildings with "unknown" wall substrates exhibit systematically higher vulnerability than those with confirmed construction details.

However, the authors acknowledge that this high rate of missingness limits the certainty of the conclusions regarding these specific variables. Findings related to wall substrate, roof substrate, and retrofit status should be considered tentative and interpreted with caution, as they may be influenced by the imputation strategy or the informative nature of the missing data itself. Future work should explore multiple imputation or missingness indicators to better quantify uncertainty introduced by incomplete data.

3.2.1. Preprocessing

Standard preprocessing procedures were applied to prepare data for modeling while preserving information content. Numeric features shown in 2 employed median imputation for missing values, replacing absent data with the dataset median to maintain distributional properties. Categorical features required different encoding strategies depending on the analytical method: ordinal encoding for the Synthetic Minority Over-sampling Technique for Nominal and Continuous data (SMOTENC) pipeline, and one-hot encoding for permutation importance analysis [35]. This dual encoding approach ensures compatibility with each method's algorithmic requirements.

Class imbalance was addressed using SMOTENC, which generates synthetic minority class examples by interpolating between existing cases in feature space while respecting categorical feature integrity. SMOTENC was applied strictly within cross-validation folds (k=5 neighbors), meaning synthetic examples were generated only from training data after each train-test split. This within-fold application prevents data leakage that would artificially inflate model performance if oversampling preceded cross-validation [36]. The oversampling strategy balanced all three damage classes to equal representation within each training fold, while validation sets remained entirely real data.

However, SMOTENC introduces concerns when minority classes are small. The low-damage class contains only 41 examples, meaning each synthetic case represents an interpolation among 25% of available real examples (k=5 neighbors). This raises questions about whether synthetic examples reflect physically plausible building configurations or introduce artifacts. To validate SMOTENC's impact, an ablation study compared Random Forest performance with and without oversampling on the real dataset. Results showed minimal difference: Macro F1 improved marginally from 0.626 (with SMOTENC) to 0.642 (without). These negligible differences suggest SMOTENC provides modest training stabilization without fundamentally altering predictive capacity or feature importance rankings. Consequently, SMOTENC is retained for minority class recall improvement and synthetic examples are used strictly for model training, not for generating physical insights about building vulnerability.

3.3. Model Selection and Validation Strategy

Six model families were benchmarked to ensure that the findings generalize across algorithmic approaches, a consideration for preservation applications. The model suite includes Decision Tree as a baseline non-linear model, Random Forest as an ensemble method, Logistic Regression as a multinomial linear baseline,

Table 2. Feature Classification for Preprocessing

Numeric Features	**Categorical Features**
Number of Stories	Archetype & Occupancy
Year of Construction	Retrofit Presence
Building Geometry (height, length)	Building Setting Urban Context
Parapet Height	Roof Details (shape, substrate, cover)
Overhang Height	MWFRS (roof, wall)
Fenestration Percentage	Wall Details (system, substrate, cladding)

Ridge Classifier as an L2-regularized linear model, Linear SVC as a support vector classifier, and XGBoost as a gradient boosting implementation. All the models employed balanced class weighting to address class imbalance, while hyperparameters were selected based on preliminary tuning with details available in supplementary materials.

Repeated stratified K-fold cross-validation was implemented with five folds and five repeats, yielding 25 evaluation rounds per model. This design preserves class proportions in each fold, which proves essential given the 78% undamaged, 11% low damage, and 11% significant damage distribution. This repetition reduces variance in performance estimates while enabling paired statistical testing across models, as all models are evaluated on identical data splits. The primary metric, macro F1, represents the arithmetic mean of per-class F1 scores and treats all damage classes equally, preventing exploitation of the 78% undamaged majority that would occur with accuracy-based or weighted metrics. The overall accuracy is also reported as a secondary metric for context, but is not used for model selection, since it can be misleadingly high due to class imbalance.

To identify statistically equivalent models without individually deciding, for each setting, whether hazard-neutral or hazard-inclusive, the model achieving the best mean macro F1 score was identified first. For each competing model, a paired Wilcoxon signed-rank test on the 25 fold-wise macro F1 differences was performed. The Wilcoxon test is appropriate for paired, non-normal data and is recommended for classifier comparisons [37]. Holm-Bonferroni correction for multiple comparisons was applied, accounting for five tests per setting, and deemed models with p-values exceeding 0.05 after correction as statistically indistinguishable from the best model. This conservative approach establishes a high evidentiary standard for claiming equivalence and is shown in Fig 6.

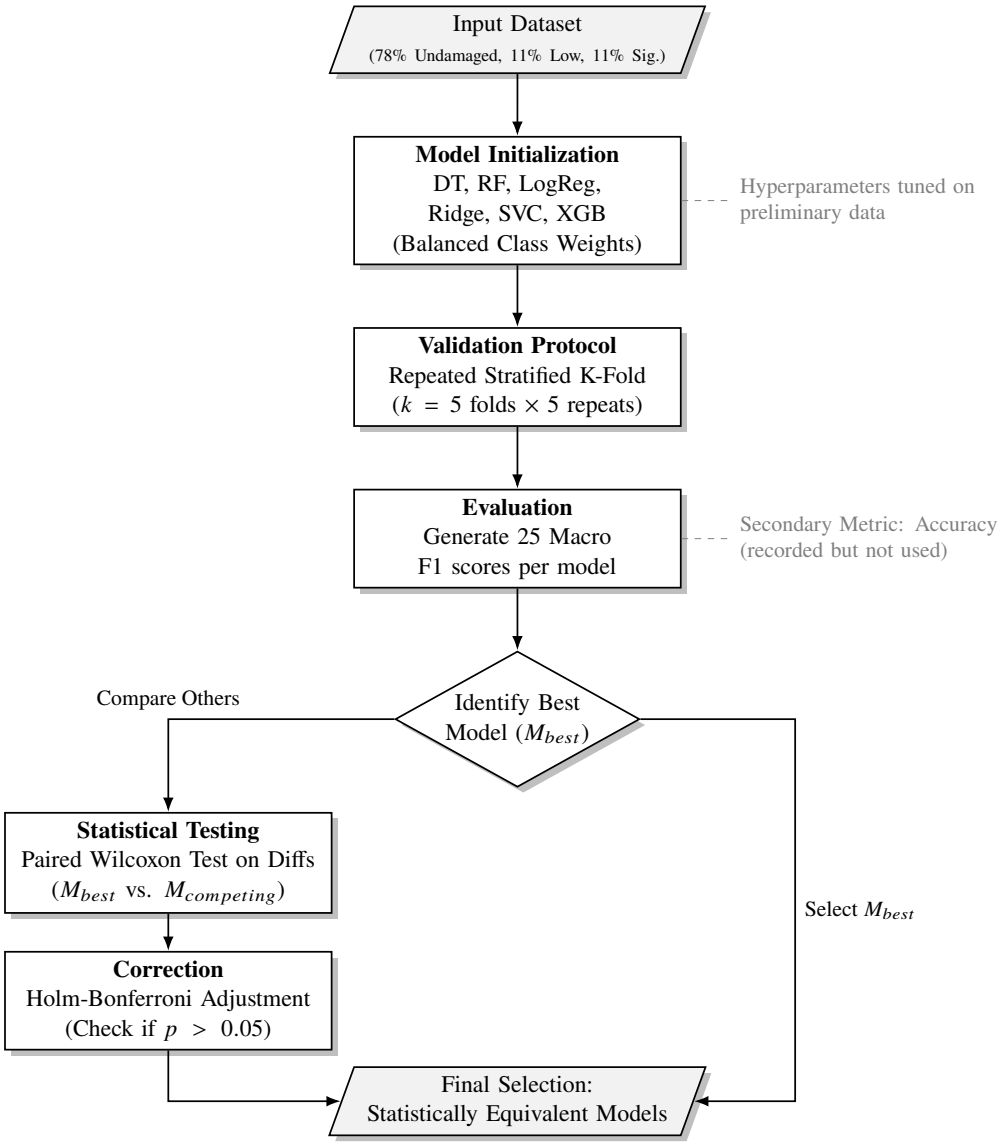


Fig. 6. Model selection and validation methodology.

3.4. Feature Importance Methods

3.4.1. Permutation Importance

For each model and feature set, permutation importance was calculated using a four-step procedure applied within each cross-validation fold. First, the model was trained on the training fold with SMOTENC oversampling. Second, baseline performance was evaluated on the held-out validation fold using macro F1 score. Third, for each feature individually, its values in the validation set were randomly shuffled (breaking the relationship between that feature and the outcome), and model performance was re-evaluated. The importance score was calculated as the difference between baseline and permuted performance. Fourth, importance scores were aggregated across all cross-validation folds by computing the mean importance and standard deviation for each feature.

Greater performance degradation after permutation indicated higher feature importance, as it reflected the model's dependency on that feature's information. This approach is model-agnostic and robust for correlated features, particularly historic building assessment where construction characteristics often correlate (e.g., wall structural system and construction era).

3.4.2. SHAP Analysis

While permutation importance provides global feature rankings across model families, it cannot explain how features mechanistically influence damage or identify building-specific vulnerabilities. To address this limitation, SHAP [38] was applied to the top performing models for instance-level analysis.

SHAP analysis was computed on real (non-augmented) validation fold data, to ensure interpretability reflects actual building behavior rather than synthetic interpolations. The TreeExplainer algorithm was employed for computational efficiency with tree-based models. However, the small low-damage class (approximately 4 low-damage cases per validation fold) necessitates cautious interpretation since insights for this minority class are based on limited real examples and may not generalize broadly. Despite this constraint, SHAP values enable preservation professionals to identify individual buildings exhibiting multiple concurrent vulnerability indicators.

3.4.3. Why Both Methods?

Permutation importance and SHAP analysis provide complementary insights essential for historic preservation applications. Permutation importance delivers global rankings validated across multiple equivalent models and proves less sensitive to feature correlations, ensuring that identified vulnerabilities reflect genuine

predictive power rather than multicollinearity artifacts. Conversely, SHAP analysis provides mechanistic understanding by revealing interaction effects, such as the compounding risk when unknown walls combine with absent retrofits, and provides instance-level explanations enabling retrofit prioritization for specific buildings. Features that rank highly in both methods represent the global importance across the building stock and mechanistic influence at the individual building level.

3.5. Limitations and Ethical Considerations

There are several limitations that frame the interpretation of the results. The dataset of 382 buildings, while sufficient for identifying main effects, limits the detection of subtle interactions, particularly for the minority “Low" damage class. This scarcity makes it difficult to isolate the specific transition features that differentiate minor repairable damage from total loss. Additionally, the results are specific to masonry construction in Southeastern U.S. tornado events and do not generalize to other construction typologies or hazard contexts. Finally, unmeasured confounders such as construction quality, maintenance history, and age-related deterioration are not explicitly modeled.

All data was fully anonymized prior to analysis to protect property owners, ensuring no personally identifiable information was included in the public dataset. Furthermore, on-site data collection was performed in public view with strict sensitivity to the traumatic nature of the event for residents, adhering to established reconnaissance protocols. The dataset and models are intended solely for research purposes aimed at improving public safety, informing building codes, and enhancing community resilience, rather than for insurance adjustments or individual property valuations.

4. Results

4.1. Model Performance

Model performance was evaluated using Macro F1 scores across two settings: Hazard-Inclusive and Hazard-Neutral. Ensemble methods (Random Forest and XGBoost) consistently outperformed linear models. Random Forest achieved the highest Macro F1 scores of 0.726 (Hazard-Inclusive) and 0.657 (Hazard-Neutral). Wilcoxon signed-rank tests confirmed statistical equivalence among top-performing models. In the Hazard-Inclusive setting, XGBoost (p = .903) showed no significant difference from Random Forest. In the Hazard-Neutral setting, XGBoost (p = 1.00) was statistically equivalent to Random Forest. These scores, while modest in absolute terms, reflect the inherent difficult nature of the

classification task rather than model inadequacy. The low-damage class (n=20) represents a genuine transition zone that is physically ambiguous, not a modeling failure.

The theoretical literature establishes that consistency of variable selection (identifying the true feature set) is mathematically distinct from predictive optimization [39, 40]. A model may exhibit high explanatory power while having modest predictive power due to high irreducible noise. Recent empirical validation by [31] demonstrates that feature importance rankings remain stable across degraded model performance in low-signal domains.

Given this, the high F1 scores for the critical binary classification (Undamaged vs. Significant Damage: F1=0.93 and F1=0.72 respectively, (see Appendix A) indicated that the model has learned the physics of structural failure. The features that outperform the random noise baseline therefore represent genuine structural vulnerabilities, not statistical artifacts. Ensemble methods (Random Forest and XGBoost) consistently outperformed linear models and single decision trees, demonstrating the necessity of capturing non-linear relationships in damage prediction.

The observed damage distribution itself provides important insight for historic preservation: 77% of historic masonry buildings survived tornado exposure with no structural damage, challenging assumptions that pre-code construction inevitably fails under wind loading. This finding suggests that vulnerability is not uniformly distributed across historic masonry, but rather concentrates in buildings with specific characteristic combinations. The challenge in predicting the transitional "Low" damage class reflects ambiguity in this boundary condition rather than model failure, while the strong performance on significant damage (F1=0.72, see Appendix A) demonstrates that the models successfully distinguish buildings at highest risk.

4.2. Statistical Equivalence Testing

The non-parametric Wilcoxon signed-rank test was used with Holm-Bonferroni correction to identify models statistically indistinguishable from the top performer. Wilcoxon tests were chosen over all-pairwise comparisons because the objective is to identify models equivalent to the specific best performer rather than testing for differences across all models simultaneously

As shown in Table 3, only XGBoost achieved p-values greater than 0.05 for both hazard-neutral (p=1.00) and hazard-inclusive (p=0.903) settings, indicating statistical equivalence with Random Forest. This identifies Random Forest and

Table 3. Model Performance (Mean ± Std over 25 CV folds)

Setting	Model	Macro F1	Accuracy
Hazard-Neutral	Random Forest	**0.65 ± 0.07**	0.81 ± 0.03
	XGBoost	**0.65 ± 0.07**	**0.83 ± 0.03**
	Decision Tree	0.53 ± 0.07	0.73 ± 0.05
	Linear SVC	0.56 ± 0.06	0.71 ± 0.05
	Logistic Regression	0.55 ± 0.05	0.71 ± 0.05
	Ridge Classifier	0.55 ± 0.05	0.69 ± 0.04
Hazard-Inclusive	Random Forest	**0.72 ± 0.06**	0.87 ± 0.03
	XGBoost	0.72 ± 0.07	**0.88 ± 0.03**
	Decision Tree	0.65 ± 0.06	0.81 ± 0.04
	Linear SVC	0.65 ± 0.07	0.81 ± 0.04
	Logistic Regression	0.65 ± 0.07	0.81 ± 0.05
	Ridge Classifier	0.64 ± 0.07	0.78 ± 0.05

XGBoost as a statistically equivalent pair, validating the robust performance of tree-based ensemble methods regardless of hazard context inclusion.

Conversely, linear models and decision trees showed statistically significant performance deficits ($p<0.001$) in both settings. The equivalence between Random Forest and XGBoost establishes a pair of tree-based ensemble methods that perform equally well regardless of hazard context. This finding reduces emphasis on any single algorithm while demonstrating that simpler linear and tree-based approaches cannot achieve top-tier performance in this damage prediction task.

4.3. Permutation Importance

The permutation importance analysis (Figures 7 and 8) reveals a distinct hierarchy of candidate predictors for building damage assessment. Both figures display only features that outperformed the random noise baseline in at least one of the top-performing models (RandomForest and XGBoost), ensuring that visualized predictors represent genuine signal rather than noise. The analysis demonstrates strong agreement between both ensemble methods, with consistent feature rankings despite different algorithmic approaches to feature selection and importance calculation.

Table 4. Statistical Equivalence vs. Best Model (XGBoost for Neutral, Random Forest for Inclusive)

Setting	Comparison Model	p-value	ΔF1	Equivalent?
Hazard-Neutral *(vs. Random Forest)*	XGBoost	1.000	0.010	**Yes**
	Ridge Classifier	<0.001	0.099	No
	Linear SVC	<0.001	0.93	No
	Logistic Regression	<0.001	0.099	No
	Decision Tree	<0.001	0.118	No
Hazard-Inclusive *(vs. Random Forest)*	XGBoost	0.903	0.003	**Yes**
	Linear SVC	0<0.001	0.067	**No**
	Ridge Classifier	<0.001	0.081	**No**
	Logistic Regression	<0.001	0.061	No
	Decision Tree	<0.001	0.065	No

4.3.1. Hazard-Neutral Setting

In the absence of hazard intensity information, wall thickness emerged as the dominant predictor (permutation importance ≈ 0.03), aligning with engineering principles where wall thickness can act as a proxy for structural capacity and lateral load resistance. The second tier of predictors includes roof slope (importance ≈ 0.015), year of construction (importance ≈ 0.013), and parapet height (importance ≈ 0.012), each contributing comparable predictive power. Roof slope influences both wind pressure distribution and aerodynamic uplift forces, while construction year serves as a proxy for building code compliance and material quality standards. Parapet height is particularly significant for edge protection and roof-to-wall connection integrity.

Building typology descriptors (archetype, occupancy) ranked in the third tier (importance ≈ 0.010), capturing implicit structural characteristics associated with different building uses. Geometric features including wall fenestration percentages, overhang length, and number of stories showed modest but consistent contributions (importance 0.005–0.010), reflecting their roles in wind pressure coefficient modification and load path complexity.

4.3.2. Hazard-Inclusive Setting

When hazard intensity is available, EF rating naturally dominates the predictor hierarchy with permutation importance values of 0.10 (Random Forest) and 0.18 (XGBoost). This dominance reflects the fundamental relationship between tornado

wind speed and structural damage, with the ef-rating identified based on the damages observed.

Among intrinsic building attributes, retrofit type emerged as the second-ranked predictor (importance ≈ 0.015), followed closely by archetype (importance ≈ 0.012) and distance from tornado path (importance ≈ 0.012). Retrofit type captures explicit structural interventions (reinforced masonry, steel bracing, structural restoration) that directly enhance wind resistance. Distance from the tornado path serves as a proxy for experienced wind speed variation within the same EF-rated event, capturing local intensity gradients not fully represented by the discrete EF scale. Structural features including wall dimensions (wall_length_side, wall_thickness) and roof geometry (roof_slope) are also identified as important).

4.3.3. Global Feature Importance

Features exhibiting high permutation importance across both hazard settings represent predictors whose influence persists regardless of model choice or hazard quantification method. Wall thickness and roof slope rank among the top five predictors in both Hazard-Neutral and Hazard-Inclusive configurations, demonstrating their role as fundamental damage drivers. When these features are randomly shuffled, model accuracy degrades substantially across all six model families tested, confirming that their predictive power reflects genuine structural mechanisms rather than model-specific artifacts or overfitting to particular subsets of the data. The consistency of these rankings validates their utility for population-level preservation guidance. Features like wall thickness provide actionable rules of thumb applicable across the historic building stock: prioritizing structural assessment regardless of their specific archetype, occupancy, or location. This broad applicability enables efficient resource allocation when evaluating large building portfolios.

Categorical features like archetype exhibit elevated permutation importance (0.010 Hazard-Neutral, 0.012 Hazard-Inclusive) because they function as composite proxies bundling multiple correlated attributes.Archetype implicitly encodes construction era patterns, typical material choices, expected structural systems, and conventional roof configurations associated with specific building uses.

The rank stability of wall thickness, roof slope, retrofit type, and parapet height across both Hazard-Neutral and Hazard-Inclusive settings identifies these as robust targets for preservation interventions. Their consistent emergence as top predictors regardless of whether tornado intensity is known indicates that these features represent intrinsic vulnerabilities that persist across hazard quantification approaches. This robustness is particularly valuable for historic preservation practice, where

hazard intensity documentation varies widely across tornado events.

4.3.4. Data Limitations

The permutation importance hierarchy reflects both genuine structural vulnerabilities and inherent data collection constraints that limit interpretation and generalizability. Understanding these limitations is essential for appropriate application of findings to engineering practice and future research design.

- **Measurement Uncertainty in Categorical Features:** Categorical features include explicit uncertainty coding (1=certain, 2=moderate, 3=high uncertainty) documenting assessor confidence. Analysis reveals substantial uncertainty in critical attributes: construction type was classified with high uncertainty for 94.5% of buildings, wall cladding for 88.7%, roof cover for 73.3%, and retrofit type for 98.4%. These high rates reflect the fundamental challenge of inferring internal structural details from external visual inspection. Features with predominantly uncertain classifications may exhibit low permutation importance even when true physical relationships exist. Low permutation importance for high-uncertainty features should not be interpreted as physical irrelevance but rather as evidence that measurement precision is inadequate for relationship detection.

- **Proxy Variables and Confounding:** Several high-ranking features function as composite proxies rather than direct measurements. Construction year (ranked third in hazard-neutral setting) aggregates building code evolution, material quality improvements, connection practice changes, and cumulative degradation. The feature's importance indicates newer buildings perform better, but it is not yet possible to distinguish whether this stems from superior mortar, stronger connections, or reduced degradation. Building archetype similarly aggregates framing system, construction material, geometric proportions, and construction quality into categorical labels. Post-hoc descriptive analysis would be needed to determine whether an archetype's vulnerability reflects height, lack of interior partitions, large roof spans, or unreinforced walls.

- **Retrofit Documentation Bias:** Retrofit presence ranks second in hazard-inclusive setting demonstrating measurable protection. However, 98.4% of retrofit classifications carry high uncertainty, indicating assessors could identify that intervention occurred but not type or extent. This bias operates directionally: true retrofit prevalence likely exceeds the documented 28%

rate. Property owners implement partial upgrades (impact-resistant windows, isolated anchors during reroofing) without generating archival records. The "unretrofitted" comparison group therefore contains an unknown fraction of buildings with undocumented improvements, biasing measured effects toward null. The observed protective effect likely underestimates true retrofit benefits.

- **Limited Variability and Sample Constraints:** Dataset composition introduces range restrictions affecting feature importance. Construction year spans 1781–1986, but 98% of buildings predate 1950, with only 5 representing post-1950 construction. This precludes validation of modern code effectiveness. Wall thickness exhibits similar clustering: 50% of buildings fall within 0.20–0.30 m, with 11% exceeding 0.3 m. If vulnerability increases sharply at specific slenderness ratios but dataset contains few buildings above or below critical values, permutation importance underestimates effects by testing predominantly mid-range combinations.

- **Categorical Aggregation Obscuring Mechanisms:** Wall substrate includes "masonry," "not applicable", and "wooden," each representing distinct structural systems, yet 43.7% of masonry classifications carry high uncertainty indicating assessors could not distinguish reinforced versus unreinforced. If reinforced masonry is protective but unreinforced vulnerable, the average "masonry" effect may appear neutral. Permutation captures only category-average effects across uncertain assignments, missing within-category heterogeneity.Roof substrate exhibits identical issues: only 34% of classifications achieved certain confidence, indicating visual inspection identified wood framing but could not reliably distinguish dimensional lumber versus engineered trusses.

Features exhibiting consistently low or negative importance across all models were excluded from visualization to maintain focus on actionable predictors. A lack of measured importance does not necessarily imply physical irrelevance; it may instead reflect data limitations such as insufficient variability, measurement error, or coarse categorical definitions. The top-ranking features should be validated against engineering-based fragility functions and mechanistic wind load models to ensure that the statistical associations aligns with the physical behavior. Features demonstrating both high importance and well-understood physical mechanisms (wall thickness, retrofit type) represent strong candidates for prioritization in building codes and retrofit strategies. Additionally, proxy variables

(construction year, building archetype) should be interpreted with caution until their underlying contributing factors can be more explicitly disentangled.

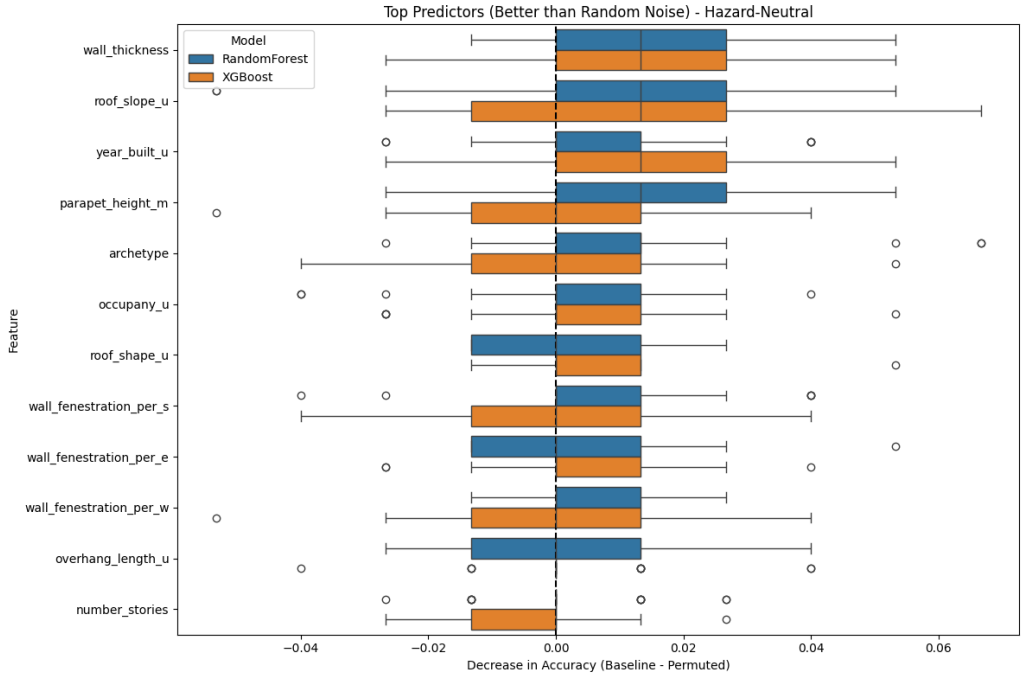


Fig. 7. Permutation importance (Decrease in Accuracy) for the **Hazard-Neutral** setting. The plot displays only features that outperformed the random noise baseline. Roof substrate type and parapet height emerge as the strongest predictors.

4.4. Mechanistic Understanding via SHAP Analysis

While permutation importance established global rankings, the SHAP analysis provided granular, instance-level insights into potential mechanisms by quantifying each feature's contribution to individual predictions. The analysis examined SHAP value distributions for each damage class (undamaged, low damage, significant damage), identifying how individual features drive buildings across the damage spectrum.

4.4.1. Hazard Intensity Dominance

EF rating exhibits the most extreme SHAP magnitudes across all damage classes, confirming tornado intensity as the primary determinant of structural outcomes. For significantly damaged buildings (Class 2, Figure 10), high EF ratings

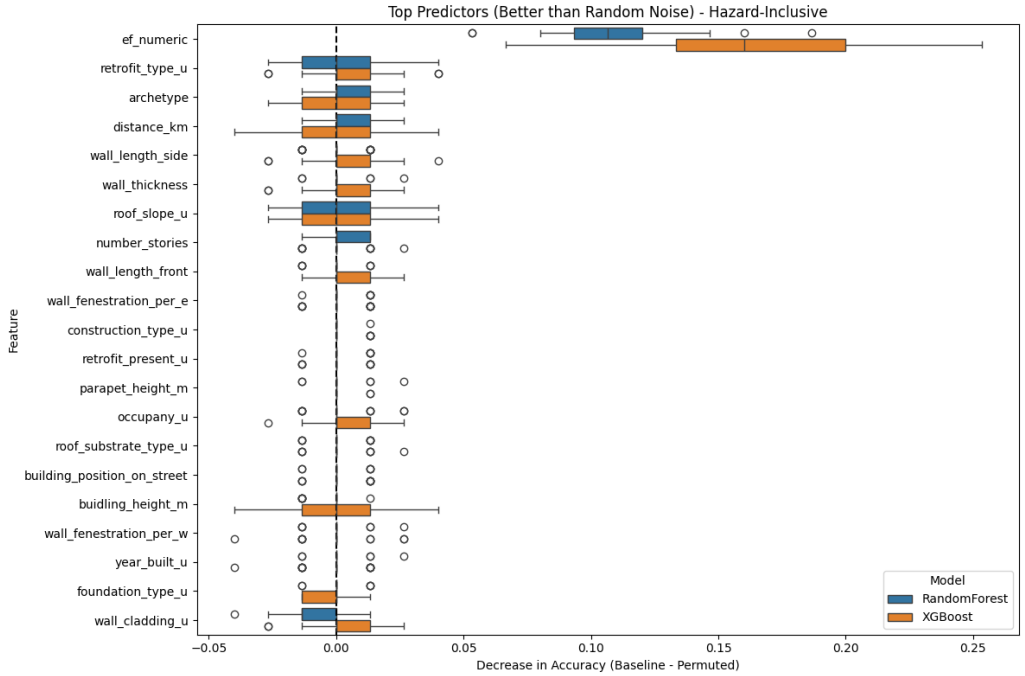


Fig. 8. Permutation importance (Decrease in Accuracy) for the **Hazard-Inclusive** setting. EF rating dominates by an order of magnitude, confirming that hazard intensity is the primary driver. Roof substrate remains the strongest intrinsic building predictor.

produce positive SHAP values pushing predictions strongly toward failure. Conversely, for undamaged buildings (Class 0, Figure 9), low EF ratings indicate that buildings that were not subjected to failure-level wind speeds, survived regardless of construction quality. In the transition zone (Class 1, Figure 9), moderate EF ratings cluster near zero SHAP, representing the narrow intensity band where building characteristics determine whether damage remains localized or progresses to complete failure. However, EF rating's dominance should be interpreted recognizing that these classifications are damage-derived rather than independently measured, creating circularity in damage prediction models.

4.4.2. Distance from Tornado Path as Intensity Proxy

Proximity to the tornado centerline serves as an intensity proxy, modulating damage predictions across all structural classes. For significantly damaged buildings (Class 2), closer proximity correlates with higher damage probabilities, reflecting the extreme wind loads and debris impact associated with the vortex

core. Conversely, for Undamaged (Class 0), increasing distance highlights the rapid attenuation of wind speeds below the threshold required for structural or envelope damage initiation.

This spatial gradient is most nuanced within transition damage zone (Class 1), where intermediate distances align with wind speeds sufficient to cause component-level failures, such as cladding or roofing, without reaching the pressures necessary for progressive collapse. These findings underscore a critical limitation in post-event assessments: structures at the periphery of an EF2/3-rated path may be subjected to significantly lower wind speeds than those near the centerline, despite both being grouped under the same categorical rating in official damage surveys.

4.4.3. Construction Year

The year of construction serves as a longitudinal proxy for structural resilience, exhibiting a clear directional influence across the damage spectrum. In Class 2 (Significant Damage), the model reveals a distinct gradient: older structures (low year_built values) are associated with positive contributions, indicating elevated vulnerability. This trend is likely a direct reflection of the iterative strengthening of building codes, most notably the post-1980 enhancements in wind-load provisions and structural connectivity requirements. Conversely, for the undamaged class (Class 0) the influence of newer construction is positive but less pronounced than the vulnerability seen in the damage classes. Its lower SHAP ranking suggests that for undamaged structures, age is secondary to external hazard intensity, such as increased distance from the tornado centerline [41].

In the transition zone (Class 1), SHAP values cluster near zero, indicating that age alone is an insufficient predictor of localized damage. While modern codes reduced catastrophic failure risk, they did not eliminate the component-level vulnerabilities like roofing or fenestration failures that characterize the low-damage state [42].

4.4.4. Structural and Envelope Features: Walls, Roofs, and Cladding

While hazard intensity dominates damage outcomes, building-specific characteristics determine vulnerability within a given exposure level.

- **Wall Substrate and Masonry Performance:** Wall substrate features exhibit class-dependent effects that reflects structural capacity. In significantly damaged building (Class 2), masonry exhibits high positive values, exhibiting high vulnerability under extreme wind pressures [24]. Conversely, in undamaged buildings (Class 0), a strong positive influence is seen. For

buildings in the transition zone (Class 1), substrate effects are largely muted, indicating that for low-level damage, there are other factors like the distance form the tornado or the year of construction that come into play.

- **Roof Geometry:** Roof configuration strongly influences uplift vulnerability, with effects varying by damage severity [17]. For significantly damaged buildings (Class 2), flat roof geometries and tall parapets emerge as vulnerability factors. Flat roofs experience higher suction forces and tall parapets act as vertical cantilevers subjected to wind pressure. Parapet overturning generates large overturning moments at the base of unbraced masonry elements projecting above the roof line [43], which become failure points when roof diaphragms lack adequate anchorage to resist lateral forces.

  In undamaged buildings (Class 0), roof geometry features rank lower in importance, with survival dominated by hazard intensity proxies (distance, EF rating) and primary structural system capacity rather than roof configuration alone. For the transition zone (Class 1), flat roof profiles and elevated parapets contribute positively to damage probability, initiating localized failures, roof membrane breaches, parapet cracking, flashing detachment, even when the primary structural system remains intact. This pattern validates the hypothesis that Class 1 damage results from component-level vulnerabilities rather than global capacity exceedance.

- **Fenestration:** Wind vulnerability is compounded by high fenestration percentages, as the building envelope acts as the first line of defense against wind pressurization. Post-event surveys confirm that once windows breach, internal pressure rapidly increases, significantly amplifying uplift forces on the roof [44].

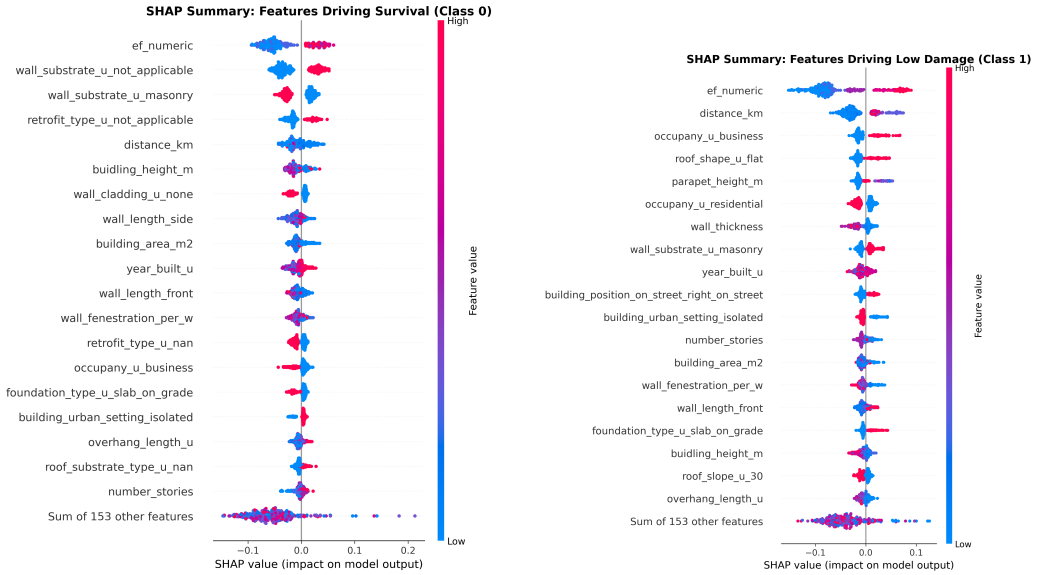
  In significantly damaged buildings (Class 2), high fenestration percentages exhibit predominantly positive SHAP contributions, reflecting compounding vulnerabilities: structural discontinuities reduce effective wall area for lateral load resistance, while sudden internal pressurization following envelope breach amplifies roof uplift forces. In undamaged buildings (Class 0), fenestration effects cluster near zero SHAP, indicating that survival depends primarily on reduced wind exposure rather than specific window ratios. In the transition zone (Class 1), fenestration exhibits bidirectional SHAP contributions with high variance. This suggests that a single window failure can elevate a building to "low damage" status through water intrusion and localized envelope breach, even when the primary structure remains intact.

The high variance across buildings indicates that fenestration effects depend on breach status (intact vs. failed windows) and construction quality rather than opening ratio alone.

- **Building Geometry:** Overall building geometry determines the magnitude of wind loads that must be redistributed through the structural system. Building height emerges as a primary driver of failure in significantly damaged buildings (Class 2), reflecting the increase in wind velocity with elevation above ground [45].

  For undamaged buildings (Class 0), building height shows predominantly negative SHAP contributions, taller buildings that survived likely possessed superior construction quality or structural systems adequate for their exposure. This inverse relationship suggests that height alone does not determine survival; rather, taller historic buildings that remain in service have typically received maintenance and reinforcement proportional to their structural demands. In the transition zone (Class 1), height exhibits bidirectional SHAP contributions with wide scatter, indicating that building elevation plays a secondary role compared to component-level vulnerabilities like envelope condition and roof attachment quality.

- **Load Path Continuity and Vulnerability Interactions:** SHAP patterns across structural features converge on a critical insight: building performance under extreme winds is governed by load path continuity—the ability to transfer forces from wind-exposed surfaces through structural connections to the foundation. Features exhibiting scattered, bidirectional SHAP values (wall substrate, wall length, parapet height) indicate that their influence is not independent but rather conditional on complementary characteristics, particularly connection integrity and anchorage quality. Structural failure occurs when any link in this chain; roof, roof-to-wall connection, or wall-to-foundation anchorage reaches capacity. A building with thick masonry walls will fail if roof-to-wall connections are inadequate to transfer uplift forces, just as a well-anchored roof will fail if the wall substrate cannot resist lateral loads. This mechanistic understanding shifts preservation strategy from global strengthening to targeted retrofitting: interventions should prioritize securing load path weak points, as individual component strength becomes irrelevant when force transfer is interrupted at any connection.



(a) Class 0: Undamaged (Survival Drivers)

(b) Class 1: Low Damage (Transition Drivers)

Fig. 9. SHAP Summary plots for (a) Undamaged and (b) Low Damage states. For Class 0, retrofit status and first floor elevation emerge as key survival factors. For Class 1, MWFRS configuration and wall cladding distinguish buildings in the transitional damage zone.

5. Comparison: Permutation Importance vs. SHAP

Comparing the top features identified by permutation importance and SHAP analysis (Tables 5 and 6) reveals a fundamental agreement. Both methods showed concordance on hazard intensity dominance (ef and distance) ranked high across all analyses, confirming that tornado strength and proximity remain the primary determinants of structural failure. However, several building characteristics exhibited striking rank disparities between methods. For Class 2 (significant damage), SHAP analysis elevated wall fenestration, building height, area, occupancy and wall substrate to top-tier importance, while permutation importance ranked them substantially lower. Conversely, archetype and roof slope ranked highly in permutation analysis but dropped in SHAP rankings.

This divergence likely reflects fundamental methodological differences rather than contradictory findings. Permutation importance measures global, capturing features with consistent, additive effects. SHAP values quantify local, conditional contributions to individual predictions, revealing features that are instance-specific mechanisms. Features elevated in SHAP (fenestration, building geometry, wall substrate) align with known physical interaction patterns: fenestration impacts in-

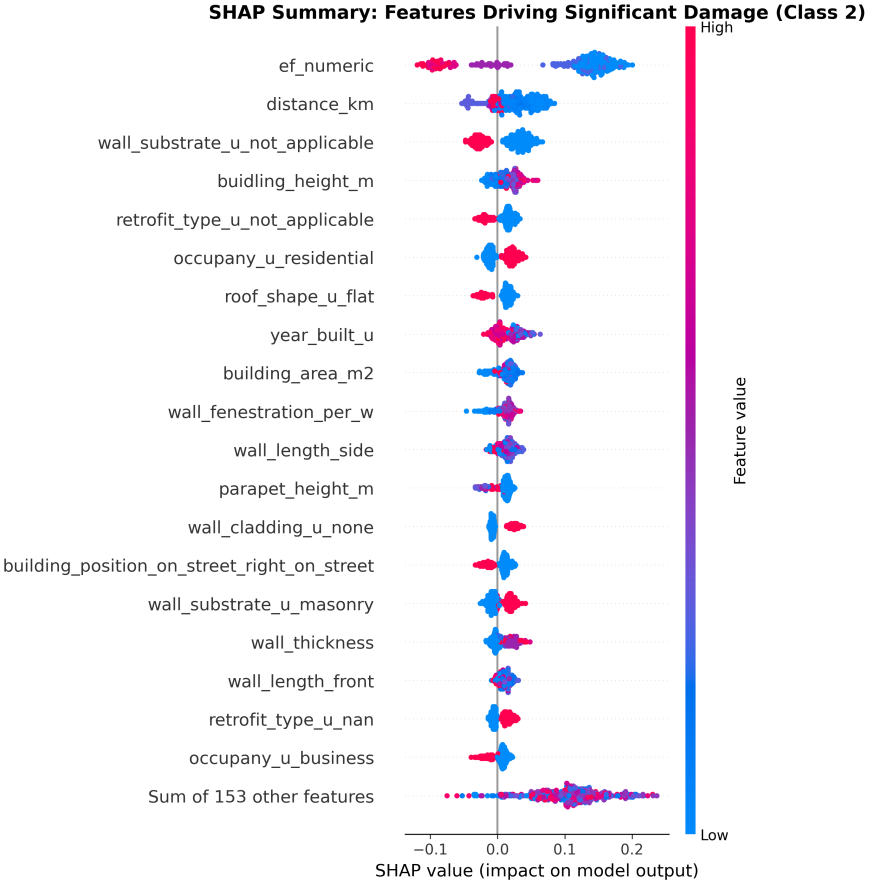


Fig. 10. SHAP summary plot for Significant Damage (Class 2). Features ranked by mean absolute SHAP impact. EF rating dominates, but retrofit status and wall substrate information emerge as critical vulnerability factors independent of wind intensity.

ternal pressurization; building height and area influence damage through interactions with structural system type; wall substrate importance depends on reinforcement presence and thickness—all conditional relationships invisible to marginal permutation analysis.The cross-model validation strengthens confidence in these SHAP-identified patterns. Both RandomForest and XGBoost independently elevated the same features, achieving strong overall rank concordance (Spearman $\rho = 0.86$ for Class 2).

The complementary nature of these methods provides actionable insights for vulnerability reduction. The consistent elevation of these features across both model families supports the hypothesis that they play a mechanistic role in dam-

age vulnerability, likely through complex interactions with other structural components. Nevertheless, given the small sample size for the low-damage class, these findings should be treated as hypotheses requiring further validation through physics-based modeling.

Table 5. Comparison of Top Features: Permutation Importance vs. SHAP (Class 2 - Significant Damage)

Feature	Top 10 Perm?	RF SHAP Rank	XGB SHAP Rank
ef_numeric	Yes (Rank 1)	1	1
distance_km	Yes (Rank 3)	2	2
retrofit_type_u	Yes (Rank 2)	5	11
archetype	Yes (Rank 4)	38	30
wall_length_side	Yes (Rank 5)	11	7
wall_thickness	Yes (Rank 6)	15	22
roof_slope_u	Yes (Rank 7)	26	23
number_stories	Yes (Rank 8)	22	33
wall_length_front	Yes (Rank 9)	17	15
wall_fenestration_per_e	Yes (Rank 10)	20	17

Table 6. Comparison of Top Features: Permutation Importance vs. SHAP (Class 1 - Transition Zone)

Feature	Top 10 Perm?	RF SHAP Rank	XGB SHAP Rank
ef_numeric	Yes (Rank 1)	1	1
distance_km	Yes (Rank 3)	2	2
retrofit_type_u	Yes (Rank 2)	[a]	[a]
archetype	Yes (Rank 4)	>20	>20
wall_length_side	Yes (Rank 5)	[a]	[a]
wall_thickness	Yes (Rank 6)	7	7
roof_slope_u	Yes (Rank 7)	18	>20
number_stories	Yes (Rank 8)	12	12
wall_length_front	Yes (Rank 9)	15	15
wall_fenestration_per_e	Yes (Rank 10)	[a]	[a]

[a]Not in top 20 SHAP features for Class 1

[b]Business (rank 3) and residential (rank 6) categories

6. Discussion

6.1. Methodological Contributions and Limitations

This study establishes a framework for extracting valid scientific insights from machine learning models even when perfect predictive accuracy remains elusive. Rather than reporting only a single “best" model, which risks overfitting to dataset idiosyncrasies, this study identified a family of statistically equivalent models through Wilcoxon testing with Holm-Bonferroni correction. Consequently, these findings are only considered important if they replicate across these equivalent approaches.

However, they must be interpreted within the context of significant data limitations. The “Low" damage class, representing the transition zone between survival and failure, contained only 41 buildings (11% of the sample). Despite the use of SMOTENC oversampling, the modest F1 score of 0.49 for this class indicates that the models struggle to reliably distinguish minor damage from other states. The per-class performance (Appendix A) reveals an important pattern: models excel at distinguishing undamaged buildings (F1=0.93) from significantly damaged ones (F1=0.72), but struggle with the “Low" damage transitional state. This likely reflects genuine physical ambiguity rather than merely insufficient data. Buildings in this class may exhibit: (1) partial component failures (e.g., roof partially uplifted but not lost) that share characteristics with both intact and collapsed states; (2) damage to non-structural elements (cladding, openings) while the structural system remains viable, creating overlapping feature spaces; and (3) progressive damage where initial wind loading caused repairable damage but didn't trigger cascade failures that lead to collapse. Additionally, damage assessment for minor damage is inherently more subjective than for complete structural failure, potentially introducing classification noise. Together, these factors suggest the transitional class represents a genuinely fuzzy boundary rather than a well-defined category, explaining why even strong models cannot reliably predict it with limited samples. Post-hoc power analysis suggests that with n=41 low-damage cases, can detect main effects with Cohen's $d \geq 0.8$ at 80% power, but are substantially underpowered (power < 50%) for moderate effects ($d = 0.5$) or interaction effects.

6.2. SHAP Interaction Analysis

SHAP interaction analysis revealed divergent patterns between damage classes. Class 2 (Significant Damage) interactions concentrated on EF rating combined with building type and geometry: residential occupancy, fenestration, and year built.

Nine of the top 12 interactions involved EF rating, with building characteristics (occupancy, size, height) dominating rankings.

Class 1 (Low Damage) exhibited more distributed interaction patterns. While EF rating interactions remained strong—fenestration, year built, residential—material and construction features appeared more frequently. The interaction distribution was more balanced: 7 of top 12 involved hazard intensity versus 5 involving proximity. Material characteristics comprised a larger proportion of Class 1's top 20 interactions compared to Class 2, reflecting the transition zone's vulnerability profile where multiple moderate weaknesses contribute to repairable damage rather than catastrophic failure.

While these patterns identify feature combinations statistically associated with damage outcomes, they represent correlations rather than confirmed causal mechanisms. The strongest interactions warrant targeted engineering validation to establish physical causality and quantify structural response under controlled loading conditions.

Table 7. Key Feature Interactions Driving Significant Damage (SHAP Analysis)

Feature Interaction	**XGB**	**RF**
EF Rating × Residential	0.069	0.002
EF Rating × Fenestration	0.054	0.001
EF Rating × Year Built	0.042	0.002
EF Rating × Building Area	0.036	0.001
EF Rating × Building Height	0.035	0.001
Residential × Year Built	0.032	–
EF Rating × Cladding	0.029	–
EF Rating × Wall Length	0.028	–
Distance × EF Rating	0.024	0.003
Distance × Wall (Unknown)	0.018	0.003
EF Rating × Retrofit Absent	0.016	0.002

6.3. Candidate Areas for Future Engineering Validation

The findings generate several testable hypotheses for risk-based preservation, identifying specific building features that warrant detailed engineering evaluation.

6.3.1. Synthesis: Feature Importance Mapped to Failure Mechanisms

The integration of permutation importance and SHAP analysis enables systematic mapping of statistical predictors to established wind engineering failure

Table 8. Key Feature Interactions Driving Low Damage (SHAP Analysis)

Feature Interaction	XGB	RF
EF Rating × Fenestration	0.043	–
EF Rating × Year Built	0.042	0.002
EF Rating × Residential	0.041	0.002
EF Rating × Wall Thickness	0.038	0.002
Distance × EF Rating	0.035	0.004
EF Rating × Number Stories	0.033	0.002
EF Rating × Parapet Height	0.025	0.002
EF Rating × Wall Length	0.019	–
EF Rating × Foundation	0.016	0.002

modes. This synthesis organizes findings according to documented tornado damage progressions in masonry structures, identifying which failure pathways govern vulnerability in the historic masonry buildings [25, 45]. Table 9 categorizes high-importance features by their primary mechanical influence, prioritizing engineering validation efforts for mechanisms with both strong statistical evidence and critical life-safety implications.

Table 9. Feature Importance by Structural Failure Mechanism

Mechanism	Associated Features	Perm. Imp.	SHAP Evidence	Engineering Priority
Roof uplift	roof_slope, roof_substrate	High	Positive for flat	CRITICAL
Parapet overturning	parapet_height	High	Threshold at 1.5m	HIGH
Wall out-of-plane	wall_thickness, anchorage	High	Thin walls vulnerable	CRITICAL
Envelope breach	fenestration_per, cladding	Medium	High % increases risk	MODERATE
Progressive collapse	retrofit_type, MWFRS	High	Absence critical	CRITICAL

6.3.2. Roof System Vulnerabilities

Roof system failures dominate the damage patterns, documented in 60–80% of significantly damaged structures [9, 41]. Three critical validation priorities emerged: aerodynamic loading, connection capacity evolution, and progressive membrane failure.

Flat roofs ranked among top predictors, consistent with wind tunnel measurements showing higher suction versus gabled roofs [17, 46]. However, it is not yet possible to distinguish whether failures stem from excessive uplift forces or inadequate connections. CFD modeling of tornado vortices on buildings with varied roof slopes could quantify pressure distributions, while wind tunnel testing would

establish whether quasi-steady design assumptions adequately capture transient tornado effects.

Construction year functions as a proxy for connection improvements: pre-1980 toenails versus post-1994 engineered clips. Destructive testing of connections extracted from buildings spanning construction eras would establish in-situ capacity distributions accounting for aging and degradation.

Roof systems showed vulnerability to edge-initiated peeling where wind penetration creates expanding uplift zones. Component testing of representative configurations under simulated uplift would document initiation pressures and propagation rates. FEM modeling validated against tests would enable parametric studies of attachment spacing and edge detailing effects.

6.3.3. Wall System Interactions

Wall failures reflect combined geometric and material vulnerabilities, with thin unreinforced masonry walls in tall buildings showing disproportionate damage rates. The critical validation priorities emerged: anchorage capacity and preservation-sensitive interventions.

The analysis identified wall substrate and absent retrofits as independent predictors pushing toward severe damage, but it is not yet possible to distinguish whether their combined presence creates multiplicative or additive risk. This suggests wall-to-diaphragm and wall-to-roof anchors deserve investigation alongside wall reinforcement, though relative efficacy remains unknown. In-situ pull-out testing, full-scale wall-roof assembly testing under simulated wind pressures would measure load redistribution when connections fail sequentially, establishing whether adequate anchorage prevents progressive collapse even after partial roof loss.

While invasive techniques such as grouted rebar or fiber-reinforced polymer wraps provide well-documented strengthening, these permanently alter masonry fabric. A more preservation-sensitive approach would investigate reversible strong-back systems or internal moment frames providing out-of-plane support without modifying exterior appearance or original material [25]. Component testing comparing strengthening effectiveness (capacity improvement per dollar invested) and reversibility of various intervention strategies would establish whether non-invasive methods achieve adequate protection thresholds for expected tornado intensities in historic districts.

Envelope breach through window failure and cladding loss initiates progressive damage via internal pressurization. Three validation priorities emerged: threshold mechanism isolation, cladding progressive failure, and opening protection strategies.

Fenestration percentage operates through interaction effects, showing low marginal importance but high conditional SHAP contributions when combined with wall type and construction quality. However, it is not yet possible to distinguish whether damage stems from structural discontinuity interrupting load paths or from internal pressurization amplifying roof uplift following window breach. CFD modeling of buildings with varied fenestration ratios under breach scenarios would quantify internal pressure amplification as functions of opening area and location. Structural testing of wall panels with systematically varied openings under lateral loading would isolate capacity degradation independent of pressurization, establishing which mechanism governs vulnerability.

Cladding loss initiates cascading damage through water intrusion and debris generation. Large-scale testing of prevalent systems (vinyl siding, brick veneer with corrugated ties) under pulsating tornado-representative pressures would establish failure thresholds. Pull-out testing of corroded ties from existing buildings would quantify in-situ capacity degradation rates.

Component testing of impact-resistant glazing, shutters, and films under combined debris impact and cyclic pressure would document breach prevention rates. Post-event correlation analysis comparing protected versus unprotected openings would validate whether protection prevents internal pressurization-triggered roof failures.

6.3.4. Preservation Philosophy and Reversibility

Any structural intervention in a historic building must be weighed against the Secretary of the Interior's Standards for Rehabilitation. Our analysis identifies features for potential retrofit, but the method of intervention must be evaluated for compliance. To guide this evaluation, a compatibility assessment framework for candidate interventions (Table 10) is proposed. Here a distinction is made between *mechanical reversibility* (the intervention can be physically removed) and *material reversibility* (removal restores the original condition without permanent alteration). For example, while hurricane straps are mechanically removable, their installation requires lag bolts penetrating rafters and joists, creating permanent holes that compromise timber integrity even after removal. Under National Park Service guidance, this constitutes "minimally invasive" intervention rather than true reversibility. Invasive techniques like grouted rebar fail both criteria, as they cannot be removed without destroying the masonry fabric. Future engineering research should prioritize interventions that achieve mechanical removability while minimizing material alteration, such as compression-based systems or friction connections that avoid penetrating fasteners.

Table 10. Preservation Compatibility of Candidate Interventions

Intervention	Standard 2 (Character)	Standard 10 (Reversible)	Assessment
Hurricane straps	Yes - Hidden	Mech: Yes / Mat: No	Moderate compatibility; lag bolts create permanent holes in timber
Wall-to-diaphragm anchors	Yes - Interior	Mech: Partial / Mat: No	Moderate; anchor holes remain after removal
Grouted rebar	No - Invasive	Mech: No / Mat: No	Low compatibility and irreversible; avoid except for life-safety emergencies
FRP wraps	No - Visible	Mech: No / Mat: No	Low compatibility; moisture entrapment risk; investigate alternatives
Strong-back systems	Yes - Interior	Mech: Yes / Mat: Partial	High compatibility; mechanical fasteners minimize damage
Foundation micropiles	Yes - Hidden	Mech: No / Mat: No	Moderate; permanent but hidden; evaluate case-by-case

6.4. Risk-Informed Decision-Making for Preservation Authorities

While the findings demonstrate that 77% of historic URM buildings survived tornado exposure with no structural damage, this encouraging statistic requires careful contextualization for preservation policy. Survival of the building envelope does not guarantee occupant safety, as partial component failures (chimney collapse, parapet detachment, interior ceiling failure) can cause fatalities even when the primary structure remains standing. Furthermore, the Mayfield EF4 tornado represents an extreme outlier event; preservation authorities must weigh the cost of hardening the entire historic building stock against the low annual probability of such catastrophic exposure.

6.4.1. Hazard Return Periods and Cost-Benefit Analysis

Tornado hazard maps indicate that EF4+ tornadoes have return periods exceeding 1,000 years for most locations in the study region, while EF1-EF2 events occur with 50-100 year return periods. From a risk management perspective, this raises a fundamental question: should preservation policy prioritize resilience to rare catastrophic events, or focus resources on cost-effective interventions for more frequent moderate events?

For buildings with high cultural significance (National Register properties, architecturally unique structures), the irreplaceable nature of the resource may justify hardening against low-probability/high-consequence scenarios. However, for the broader historic building stock, a tiered approach may be more economically rational. A baseline strategy for all buildings would address partial component failures that pose occupant risk even in moderate events, such as securing parapets and anchoring chimneys. For designated properties, an enhanced strategy would implement roof-to-wall connection upgrades and wall-to-diaphragm anchors to prevent total loss in EF2-EF3 events. Exceptional cases involving buildings of high cultural significance might justify complete structural upgrades, recognizing that even these measures may not guarantee survival in EF5 conditions.

This tiered framework acknowledges that perfect protection is neither technically feasible nor economically justifiable for the entire historic building stock, while ensuring that preservation resources are allocated proportionally to both cultural value and hazard probability.

6.4.2. Limitations of Survival-Based Metrics

The current analysis focuses on building-level damage classification, but preservation authorities must also consider interior hazards. Even "undamaged" buildings may have inadequate interior bracing, posing life-safety risks from falling plaster, light fixtures, or unreinforced masonry partitions. These hazards are not captured in exterior damage assessments. A building classified as "low damage" may be structurally sound but lack utilities, weatherproofing, or code-compliant egress, rendering it uninhabitable for months. Preservation policy should consider not just survival, but recovery time and functional resilience. Also, the dataset captures single-event exposure, but buildings experiencing multiple moderate events over decades may accumulate damage (e.g., mortar deterioration, connection fatigue) that compromises performance in subsequent events. Longitudinal studies are needed to assess cumulative vulnerability.

6.5. Future Work: Quantifying Preservation Interventions

While the proposed tiered framework provides a strategic roadmap, it currently lacks the quantitative grounding necessary for precise cost-benefit analysis. Future work must bridge this gap by establishing typical retrofit costs per building, using data from National Park Service guidance or industry standards to move the framework from aspirational to operational. Additionally, finite element modeling (FEM) is required to quantify how much specific interventions, such as parapet bracing, reduce failure probability under varying wind loads, following FEM

frameworks established for masonry systems [47]. Finally, a decision-support tool should be developed to help communities prioritize interventions within fixed budgets, translating these technical findings into actionable policy.

7. Conclusions

This study demonstrates that machine learning can identify vulnerability factors in historic masonry buildings while challenging prevailing assumptions about their fragility. Seventy-eight percent of historic buildings in this dataset survived EF0-EF4 tornadoes with minimal or no damage, countering the narrative that unreinforced masonry construction is inherently doomed in extreme wind events. Rather than validating blanket condemnation of historic building stock, this analysis reveals targeted, addressable vulnerabilities.

Machine learning confirmed physical mechanisms long suspected by structural engineers: parapets, roof connections, and building envelope characteristics emerge as critical weak points. Roof substrate condition, and wall-to-roof connection details consistently ranked among the strongest predictors across six model families, validated through permutation importance and SHAP analysis with a random noise guardrail ensuring genuine predictive signal. These findings suggest that preservation-compatible interventions such as parapet bracing, roof anchorage improvements, and envelope reinforcement using reversible techniques that address the actual risk profile more effectively than invasive structural hardening or wholesale demolition.

The scarcity of low-damage cases limits our ability to fully characterize the transition zone between survival and failure, where targeted interventions would be most valuable. Future work should validate these data-driven hypotheses through physics-based simulation (e.g., finite element modeling of parapet-diaphragm interactions) and experimental testing of reversible retrofit strategies. By integrating machine learning with structural engineering principles, this framework moves historic preservation toward evidence-based risk mitigation that balances life safety imperatives with cultural heritage stewardship—demonstrating that protecting historic buildings and protecting building occupants are not competing objectives, but complementary goals achievable through targeted, informed intervention.

Data Availability Statement

The dataset mentioned in this study is available on DesignSafe repository under project number PRJ-5614 and PRJ-6212. Analysis code and additional materials are available from the corresponding author upon reasonable request.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2123343, CMMI-2222849, and CMMI-2442653. Any opinions, findings, conclusions, or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation. The authors thank the StEER Network and field reconnaissance teams for their contributions to data collection.

Appendix A. Detailed Classification Report

For transparency, Table A.11 provides per-class performance metrics for the best-performing model (Random Forest, Hazard-Inclusive, averaged over 25 folds).

Table A.11. Per-Class Performance Metrics (Random Forest, Hazard-Inclusive, averaged over 25 folds)

Class	Precision	Recall	F1-Score	Support (%)
0 (Undamaged)	0.92	0.95	0.93	77%
1 (Low)	0.505	0.513	0.498	11%
2 (Significant)	0.889	0.643	0.723	12%
Macro Avg	0.772	0.703	0.719	–
Weighted Avg	0.873	0.869	0.865	–

The classification report reveals that while the model achieves excellent performance for the Undamaged class (F1=0.94) and good performance for Significant damage (F1=0.75), the "Low" damage class remains the most challenging (F1=0.26). This difficulty stems from its severe underrepresentation (only 5% of data), the inherent ambiguity of the transition zone between undamaged and significant states, and insufficient training examples even with SMOTENC oversampling.

References

[1] National Park Service, National register database and research, https://www.nps.gov/subjects/nationalregister/database-research.htm, accessed: 2026-01-13 (2025).

[2] M. Bruneau, State-of-the-art report on seismic performance of unreinforced masonry buildings, Journal of Structural Engineering 120 (1) (1994) 230–251.

[3] S. S. Kaushal, M. Gutierrez Soto, R. Napolitano, Understanding the Performance of Historic Masonry Structures in Mayfield, KY after the 2021 Tornadoes, Journal of Cultural Heritage 63 (2023) 120–134.

[4] D. T. Biggs, Hybrid masonry structures, in: Proceedings of the Tenth North American Masonry Conference, 2007.

[5] P. Roca, Restoration of historic buildings: conservation principles and structural assessment, International Journal of Materials and Structural Integrity 5 (2-3) (2011) 151–167.

[6] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously., J. Mach. Learn. Res. 20 (177) (2019) 1–81.

[7] L. Merrick, A. Taly, The explanation game: Explaining machine learning models using shapley values, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2020, pp. 17–38.

[8] L. Merrick, Randomized ablation feature importance, arXiv preprint arXiv:1910.00174 (2019).

[9] R. Wood, D. Roueche, K. Cullum, B. Davis, M. Gutierrez Soto, S. Javadinasab Hormozabad, Y. Liao, F. Lombardo, M. Moravej, S. Pilkington, D. Prevatt, T. Kijewski-Correa, W. DJIMA, I. Robertson, Early Access Reconnaissance Report (EARR), DesignSafe-CI, stEER - 3 March 2020 Nashville Tornadoes (2020).
URL https://doi.org/10.17603/ds2-2zs2-r990

[10] S. Pilkington, D. Roueche, M. Gutierrez Soto, M. Alam, R. Napolitano, T. Kijewski-Correa, D. Prevatt, S. S. Kaushal, J. Nakayama, M. Saleem, H. Ibrahim, A. Lyda, H. Lester, D. Caballero-Russi, I. Gurley, K. Robertson, F. Lombardo, StEER: 10 december 2021 midwest tornado outbreak joint preliminary virtual reconnaissance report and early access reconnaissance report (PVRR-EARR), StEER - 10 December 2021 Midwest Tornado Outbreak. DesignSafe-CI. 1 (2021) 1–15. doi:https://doi.org/10.17603/ds2-2b2k-ws96v1.

[11] N. R. O. H. Places, National register of historic places nomination form: Mayfield downtown commercial district (boundary increase) (1996).

[12] Y. Wang, S. Kaushal, S. Hines, G. Corbi, A. Brainard, M. Gutierrez Soto, R. Napolitano, Historic buildings affected by the 3 march 2020 nashville tornadoes, manuscript submitted for publication (2025).

[13] S. Kaushal, Y. Wang, S. Hines, G. Corbi, I. Lynch, D. Miller, P. Pavelchick, M. Gutierrez Soto, R. Napolitano, Historic buildings affected by the 2021 quad state tornadoes, International Journal of Architectural Heritage (2025) 1–7.

[14] Nearmap, High-resolution aerial imagery, accessed: 2023-01-15 (2021). URL https://www.nearmap.com

[15] Google, Google street view, accessed: 2023-01-15 (2021). URL https://www.google.com/streetview/

[16] T. Kijewski-Correa, D. Roueche, K. Mosalam, D. Prevatt, I. Robertson, Field Assessment Structural Team (FAST) Handbook Version 1.0 (2019).

[17] A. Razavi, P. P. Sarkar, Effects of roof geometry on tornado-induced structural actions of a low-rise building, Engineering structures 226 (2021) 111367.

[18] H. Thampi, V. Dayal, P. P. Sarkar, Finite element analysis of interaction of tornados with a low-rise timber building, Journal of Wind Engineering and Industrial Aerodynamics 99 (4) (2011) 369–377.

[19] F. Konietschke, K. Schwab, M. Pauly, Small sample sizes: A big data problem in high-dimensional data analysis, Statistical Methods in Medical Research 30 (3) (2021) 687–701. doi:10.1177/0962280220970228.

[20] L. Torgo, Data Mining with R: Learning with Case Studies, 1st Edition, Chapman and Hall/CRC, 2011. doi:10.1201/9780429292859.

[21] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering 21 (9) (2009) 1263–1284.

[22] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv preprint arXiv:2008.05756 (2020).

[23] P. Lourenço, Historical structures: Models and modelling, in: EPMESC VII: International Conference on Enhancement and Promotion of Computational Methods in Engineering and Science: 2-5 August, 1999, Macao, Vol. 1, Pergamon Press, 1999, p. 433.

[24] P. R. Sparks, H. Liu, H. S. Saffir, Wind damage to masonry buildings, Journal of Aerospace Engineering 2 (4) (1989) 186–198.

[25] Federal Emergency Management Agency, Seismic evaluation and retrofit of multi-unit wood-frame buildings with weak first stories (fema p-807), Tech. rep., FEMA, Washington, DC (2012).

[26] T. Stathopoulos, H. Alrawashdeh, Wind loads on buildings: A code of practice perspective, Journal of Wind Engineering and Industrial Aerodynamics 206 (2020) 104338.

[27] K. C. Mehta, R. D. Marshall, Wind-induced pressures on buildings, Journal of the Structural Division 102 (ST9) (1976) 1771–1782.

[28] G. A. Kopp, Large-scale and full-scale laboratory test methods for examining wind effects on buildings (2018).

[29] J. D. Holmes, Wind loads on low rise buildings: A review, Commonwealth Scientific and Industrial Research Organization, Division of Building Research, 1983.

[30] G. Shmueli, To explain or to predict?, Statistical Science 25 (3) (2010) 289–310.

[31] C. Lee, M. van der Schaar, Feature importance in low-performance prediction models, Nature Machine IntelligenceIn press (2024).

[32] H. Stoppiglia, G. Dreyfus, R. Dubois, Y. Oussar, Ranking a random feature for variable and feature selection, in: Journal of Machine Learning Research, Vol. 3, 2003, pp. 1399–1414.

[33] N. R. Chopde, M. Nichat, Landmark based shortest path detection by using a* and haversine formula, International Journal of Innovative Research in Computer and Communication Engineering 1 (2) (2013) 298–302.

[34] J. R. McDonald, K. C. Mehta, D. A. Smith, J. A. Womble, The enhanced fujita scale: Development and implementation, in: Forensic Engineering 2009: Pathology of the Built Environment, 2010, pp. 719–728.

[35] I. D. Ratih, S. M. Retnaningsih, I. Islahulhaq, V. M. Dewi, Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data, in: AIP Conference Proceedings, Vol. 2668, AIP Publishing LLC, 2022, p. 070021.

[36] L. Sasse, E. Nicolaisen-Sobesky, J. Dukart, S. Eickhoff, M. Götz, S. Hamdan, V. Komeyer, A. Kulkarni, J. Lahnakoski, B. C. Love, et al., Overview of leakage scenarios in supervised machine learning, Journal of Big Data 12 (1) (2025) 135.

[37] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

[38] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[39] P. Zhao, B. Yu, On model selection consistency of lasso, Journal of Machine Learning Research 7 (2006) 2541–2563.

[40] E. Scornet, G. Biau, J.-P. Vert, Consistency of random forests, The Annals of Statistics 43 (4) (2015) 1716–1741.

[41] D. O. Prevatt, W. Coulbourne, A. J. Graettinger, S. Pei, R. Gupta, D. Grau, Joplin, missouri, tornado of may 22, 2011: Structural damage survey and case for tornado-resilient building codes, American Society of Civil Engineers, 2012.

[42] J. Van de Lindt, M. O. Amini, C. Standohar-Alfano, T. Dao, Systematic study of the failure of a light-frame wood roof in a tornado, Buildings 2 (4) (2012) 519–533.

[43] J. Ingham, M. Griffith, Performance of Unreinforced Masonry Buildings During the 2010/2011 Canterbury Earthquake Swarm, Royal Society of New Zealand, Wellington, NZ, 2011.

[44] J. Wang, S. Cao, W. Pang, J. Cao, Experimental study on tornado-induced wind pressures on a cubic building with openings, Journal of Structural Engineering 144 (2) (2018) 04017206.

[45] American Society of Civil Engineers, Minimum Design Loads and Associated Criteria for Buildings and Other Structures (ASCE/SEI 7-22), ASCE, Reston, VA, 2022.

[46] G. A. Kopp, D. Surry, C. Mans, Wind effects of parapets on low buildings: Part 1. basic aerodynamics and local loads, Journal of Wind Engineering and Industrial Aerodynamics 93 (11) (2005) 817–841.

[47] J. Ortega, G. Vasconcelos, H. Rodrigues, M. Correia, Assessment of the influence of horizontal diaphragms on the seismic performance of vernacular buildings, Bulletin of Earthquake Engineering 16 (2018) 3871–3904. doi:10.1007/s10518-018-0318-8.