Machine Learning-Based Sensitivity of Steel Frames with Highly Imbalanced and High-Dimensional Data

Hyeyoung Koh[1] and Hannah B. Blum[1,*]

[1]Department of Civil and Environmental Engineering, University of Wisconsin - Madison, Wisconsin, United States; ORCID: 0000-0002-5795-4417
[1]Department of Civil and Environmental Engineering, University of Wisconsin - Madison, Wisconsin, United States ORCID: 0000-0003-4672-0903
[*]Corresponding author: hannah.blum@wisc.edu

Abstract

The machine learning-based feature selection approach is presented to estimate the effect of uncertainties and identify failure modes of structures that incorporate a low failure probability and high-dimensional uncertainties. As structures are designed to have few failures, a dataset classified based on the failure status becomes imbalanced, which poses a challenge for the predictive modeling of machine learning classifiers. Moreover, in order to improve the accuracy and efficiency of the model performance, it is necessary to determine the critical factors and redundant factors, especially for a large feature set. This study benchmarks the novel method for sensitivity analysis by using datasets that exacerbate the problems involved in class imbalance and large input features. This study investigates two planar steel frames with spatially uncorrelated properties between structural members. Geometric and material properties are considered as uncertainties, such as material yield stress, Young's modulus, frame sway, and residual stress. Six feature importance techniques including ANOVA, mRMR, Spearman's rank, impurity-based, permutation, and SHAP are employed to measure the feature importance and identify parameters germane to the prediction of structural failures. Logistic regression and decision tree models are trained on the important feature set, and the predictive performance is evaluated. The use of the feature importance approach for structures with a low probability of failure and a large number of uncertain parameters is validated by showing identical results with the reliability-based sensitivity study and appropriate predictive accuracy.

1 Introduction

A reliability-based sensitivity analysis estimates how uncertainty in the input parameters affects system performance by analyzing the dependence of the failure probability on the inputs, which requires repeated evaluation of the performance function, resulting in significant computational cost and time. This challenge is exacerbated for large-scale engineering problems which often carry a large quantity of uncertain parameters. Researchers have made efforts to improve the computational efficiency of sensitivity analysis. For example, Wu [1] proposed the adaptive importance sampling approach, which improves computational efficiency by minimizing oversampling in the safe region of the limit-state surface. The score function approach proposed by Rubinstein and Kroese [2] estimates all sensitivities by the gradient and derivative of parameters. The score function method does not require additional simulations for reliability sensitivity analysis. Torii et al. [3] applied polynomial expansions to the performance function and its derivatives for the probability of failure sensitivity analysis. Proppe [4] introduced the local reliability sensitivity analysis using the moving particles method, which estimates the failure probability based on the new locations of the moved data points in the design space.

In recent years, the interest in artificial intelligence has been growing in the field of structural engineering because it provides efficient solutions to the problems in this field relative to traditional computational techniques. The use of artificial intelligence in steel structural design has been focused on artificial neural networks for design of steel members or connections such as compression members [5], steel panels [6], steel connections [7, 8], and cold-formed steel channels [9]. For reinforced concrete (RC) members or systems, machine learning techniques are implemented to predict structural responses such as the shear capacity of fiber RC beams [10], structural response of RC deep beams [11] and RC slabs [12]. Data-driven

machine learning approaches were used for fragility, risk, and vulnerability assessment of a special steel moment resisting frame building [13] and RC building frames [14]. Also, machine learning approaches were used to identify failure modes and rank the significant factors affecting the failure mode of RC members [15], RC frames [16], and steel frames [17]. Regarding the assessment of structural members of systems using machine learning techniques, RC structures have been mainly investigated rather than steel structures. In particular, studies on steel structures have used neural networks, which is a specific technique of machine learning, and have focused more on structural members than systems.

Koh and Blum [17] introduced the machine learning-based feature selection framework to be used for structural sensitivity analysis. This framework measures the feature importance of all parameters and ranks them to determine the important or redundant parameters for the prediction of system failure. Two planar steel frames were investigated with the consideration of uncertainties that affect steel frame behaviors, such as yield stress, Young's modulus, frame sway, and residual stress. The frames have different failure modes and the ultimate frame strength obtained from finite element analysis was used as the response variable. The feature rankings derived by four feature importance techniques showed the identical order of factors resulting in the largest failure probability obtained from the conventional sensitivity studies. It is demonstrated that the general procedure of the proposed feature importance method can be used for sensitivity analysis. The approach is efficient because all feature importances are estimated from a single training. Moreover, variable space can be reduced by removing irrelevant parameters to improve both computational efficiency and accuracy.

This study benchmarks the feature importance approach on the datasets that incorporate a larger set of uncertainties, more data points, and a high imbalanced ratio. The importance of benchmark studies has been emphasized in the machine learning community [18, 19]. Machines carry out tasks based on learning from a given dataset, therefore the best algorithm will not be the same for all the datasets [20]. Therefore, benchmarking of the feature importance approach must be accomplished to draw conclusions for the use of the approach in a wide range of structural systems.

The steel frames investigated in this study have the same layout as for Koh and Blum [17] but have a different spatial correlation scenario. Unlike the correlated scenario that considers all columns (or beams) to have the same properties, this study applies the uncorrelated scenario where all structural members have different properties. Sensitivity studies on structural systems with uncorrelated properties inform which specific structural member largely influences the frame failure. However, the increased number of uncertainties poses additional challenges in performing reliability sensitivity studies, which require repeated evaluations of the performance function. Regarding the feature importance approach, a large set of uncertainties would make the feature importance method difficult to derive consistent rankings between various techniques. In addition, the uncorrelated scenario contains a fewer number of failures than the correlated case, therefore the class-imbalanced ratio increases. Machine learning classifiers are sensitive to imbalanced data because the prediction is biased towards the majority class [21, 22]. As structural failures should always occur rarely in structural design, it is critical to overcome the class imbalance problem. In summary, this study aims to examine how the feature importance approach works on structural sensitivity analysis when fitting high-dimensional and extremely class-imbalanced data, which presents challenges in model training for structural engineering problems.

This study implemented six existing feature importance methods to measure the importance score. There are typically two categories in feature importance methods: (1) data analysis techniques that directly analyze the data without model fitting to measure the feature importance and (2) model analysis techniques that identify important features based on predictions from trained models [23]. (1) ANOVA (Analysis of variance), (2) mRMR (minimal-redundancy-maximal-relevance) [24], and (3) Spearman's rank correlation coefficient [25] are utilized as data analysis techniques. This study used a decision tree classifier [26] to measure the feature importance by using model analysis techniques. Model analysis techniques include two feature importance methods, which are (4) impurity-based importance and (5) permutation importance, and (6) SHAP (SHapley Additive exPlanations) [27]. Based on the measured feature importance, logistic regression [28] and decision tree [26] models are fitted to predict whether a steel frame fails. The predictive performance is evaluated by specificity, recall, and the Matthews correlation coefficient [29]. The results are compared with the reliability-based sensitivity analysis results to validate the feature importance framework. Finally, the best feature importance technique depending on the failure modes is discussed.

2 Machine learning techniques for the feature importance approach

The primary task of this study is to estimate the effect of uncertainties on failure modes of steel frames using the machine learning-based feature importance approach. In a high-dimensional dataset, it is important to determine the relevant features and remove the redundant features in an effort to prevent overfitting of a model and reduce training time, thereby improving the model performance and computational efficiency. This study measures the feature importance and compares the critical features between the feature selection methods. Once the feature rankings are derived from the feature importance techniques, a classification model is employed to predict the failure status of a structure by increasing the size of the feature set used for training. The model accuracy of each feature importance method is measured to determine how many features are informative or irrelevant for identifying structural failures. The feature ranking as well as the model performance are considered when evaluating the feature importance techniques for structural sensitivity analysis.

2.1 Feature importance techniques

2.1.1 ANOVA

The ANOVA test compares the relationship between features and response variables based on the value of F-statistic. The feature importance score J_{ANOVA} is equal to the value of F-statistic. The score can be calculated by Eq. 1:

$$J_{ANOVA}(x_i) = \frac{\sum\limits_{m=1}^{M} N_m(\bar{x}_m^{(i)} - \bar{\bar{x}}^{(i)})/(M-1)}{\sum\limits_{m=1}^{M} (N_m - 1)s_m^2/(N-1)} \tag{1}$$

where N_m = the number of instances that $y = m$, $\bar{x}_m^{(i)}$ = the sample mean of feature x_i for class m, s_m^2 = the sample variance of feature x_i for class m, $\bar{\bar{x}}^{(i)}$ = the grand mean of feature x_i, M = the number of classes, M = 2 in a binary dataset, N = the total number of instances. The importance score is the ratio of between-group variance to within-group variance, thus this technique assesses the difference between the mean values of the corresponding feature x_i between the classes. A higher value of the F-statistic indicates a larger difference between the mean values among the classes, thus the feature has a significant effect on the classes. Note that ANOVA is always positive because it is based on variance, which is always positive.

2.1.2 mRMR

The mRMR technique [24] ranks features by mutual information, which considers both relevance and redundancy of features. The feature relevance indicates a correlation with the response variable, and the feature redundancy represents the information duplicated between features. As the dataset is discrete not continuous, the mutual information difference (MID) is used as the mRMR criterion, which can be estimated by Eq. 2:

$$J_{mRMR}(x_i) = I(x_i, y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_i, x_j) \tag{2}$$

where $|S|$ = the feature set size (number of features), S = a feature set, x_j = a feature not selected in the set S, and I = the mutual information. The first term represents the relevance of the feature x_i about the response variable y. The relevance is determined from the outcome variable prediction. The second term estimates the redundancy, which is measured within the selected features x_i and x_j. By intuition, a feature with a negative J_{mRMR} value has a small relevance and large redundancy, therefore, including it in model training would decrease the predictive performance of the model.

2.1.3 Spearman's rank

The Spearman's rank correlation coefficient [25] measures a monotonic nonlinear relationship between two variables, a feature x_i and the response variables y. The measured coefficient varies between -1 as the perfect negative correlation and +1 as the perfect positive correlation. A feature with the largest absolute value is considered the most important.

2.1.4 Impurity-based importance

Impurity-based importance considers the node impurity in a tree to estimate the importance. A node containing instances of one class only is pure while a node with greater than or equal to two classes is impure. The impurity-based feature importance can be computed by Eq. 3 [30], which measures the impurities at a node j before and after splitting and then averages the impurity decrease by $N^{(i)}$, which is the number of nodes in a tree split based on x_i. A negative feature importance value ($J_{impurity}$) indicates that including it in model training would decrease the predictive performance of the model.

$$J_{impurity}(x_i) = \frac{\sum\limits_{i \in N^{(i)}} (\text{Impurity before node } k - \text{Impurity after node } k)}{|N^{(i)}|} \tag{3}$$

2.1.5 Permutation importance

Permutation importance measures the importance by removing a single feature column in a dataset. First, a tree-based algorithm is fitted to obtain the baseline model performance. After training the model, a single feature column is randomly shuffled to remove the association between the feature and the response variable. The performance of the permuted model is evaluated and compared with the baseline model performance. The difference in accuracy is considered to be the importance score as shown in Eq. 4. The feature that results in the largest Mean Decrease in Accuracy (MDA) is the most important. The permutation method provides a negative score when a feature has no effect and shuffled data are shown to be more accurate.

$$\begin{aligned} J_{permutation}(x_i) = {} & \text{accuracy for dataset without permutation} \\ & - \text{accuracy for permuted dataset of } x_i \end{aligned} \tag{4}$$

2.1.6 SHAP

The SHAP algorithm [27] identifies how much each feature contributes to the response variable based on the predictions for linear models trained on all feature subsets. The difference of the predictions from the model $f_{S \cup \{i\}}$ trained on a feature subset S including a feature x_i and another model f_S excluding x_i is interpreted as the effect of x_i. The SHAP importance score is a weighted average of all possible differences, as shown in Eq. 5:

$$J_{SHAP}(x_i) = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{5}$$

where F = the set of all features, and S = all feature subsets without x_i. Since all possible subsets are used to measure the SHAP score, the computation time of the SHAP algorithm is expensive because the model is repeatedly trained on all possible feature subsets. SHAP measures the influence of features in terms of the prediction of either positive (minority class) or negative (majority class) outcomes.

2.2 Classification-based techniques

2.2.1 Logistic regression

Logistic regression [28] is trained using the top-k features obtained from the feature importance methods, where k = the number of important features. Logistic regression is used for the classification problems, which uses the logistic sigmoid function and transforms the output into a probability value between 0 and 1 as follows in Eq. 6:

$$P(Y = m) = \frac{1}{1 + exp(-(w_0 + \sum\limits_{i \in n} w_i x_i))} \tag{6}$$

where $P(Y = m)$ is the probability of presence of Class m; Class 0 represents no failure and Class 1 represents failure in this study, w_i = regression coefficients, x_i = input features, and n = the number of input features.

A linear function is embedded in the logistic regression model, which is given as the natural logarithm of the ratio of $P(Y=1)$ to $P(Y=0)$. Logistic regression estimates the regression coefficients by minimizing the value of the ratio shown in Eq. 7.

$$log(\frac{P(Y=1)}{1-P(Y=1)}) = log(\frac{P(Y=1)}{P(Y=0)}) = w_0 + \sum_{i \in n} w_i x_i \qquad (7)$$

Feature scaling was performed on the datasets used for training a logistic regression classifier, which is sensitive to the location of data points. A decision tree is scale-invariant because it trains the model based on decision rules. Therefore, the datasets for the logistic regression algorithm are transformed to a standardized scale, which has a mean value = 0 and standard deviation = 1.

2.2.2 Decision tree

A tree-based classifier, decision tree [26], is used to measure the performance in addition to estimate the feature importance score of the model analysis techniques including impurity-based, permutation, and SHAP. Decision tree continuously splits the data according to a certain parameter such as impurity or entropy. Decision rules for splitting and the leaf nodes are the final outcomes of the decision tree. The decision tree algorithm used in this study splits the nodes based on entropy (Eq. 8) until all leaves are pure:

$$H(S) = -\sum_{m=1}^{c} p_m \log_2 p_m \qquad (8)$$

where p_m = the probability of Class m occurring in the data.

2.3 Evaluation metric

After training a model with a training set, the model performance is evaluated by using a test set, which was not involved in training. A confusion matrix provides a visualization of the model performance and it is used as a performance measurement for a machine learning classification problem. The confusion matrix for binary classification consists of the four different cases as shown in Eq. 9:

$$M = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} \qquad (9)$$

where True Positive (TP) = the number of actual positives that are correctly predicted positives, True Negative (TN) = the number of actual negatives that are correctly predicted negatives, False Negative (FN) = the number of actual positives that are incorrectly predicted negatives, and False Positive (FP) = the number of actual negatives that are incorrectly predicted positives.

Several statistical rates can be computed based upon the values given in the confusion matrix. For example, Accuracy is defined as the ratio of the correctly predicted instances to all the instances, $(TP + TN)/(TP + TN + FP + FN)$. F1-score represents the harmonic mean of precision and recall, where precision is $TP/(TP + FP)$ and recall is $TP/(TP + FN)$. Accuracy and F1-score are the most popular metrics, but they lead to the overoptimistic inflated measures especially on imbalanced datasets because several classifiers learn towards the majority class [21, 22]. For instance, when a dataset has 0.01% minority class and a model predicts all minority classes incorrectly, i.e., all data points are classified as the majority class, the model accuracy is 99.99%, which is nearly perfect. In structural engineering practice, however, it is critical to identify structural failures (the minority class) rather than safe structures (the majority class), thereby necessitating using the right metric that can correctly predict both classes in a binary classification.

When a structure is designed to have little load redistribution, the target reliability index β_T is 2.7 [31], which corresponds to the failure probability P_f of 3.5×10^{-3}, derived from $\beta = -\Phi^{-1}(P_f)$, where $\Phi^{-1} =$ inverse standard normal cumulative density function. Moreover, ASCE 7 [32] suggests using a target reliability index between 3.0 and 4.0 for structural components subjected to dead, live, and other loads except earthquake loads, depending on the risk category from I through IV where category I represents the lowest level of risk to human life. Likewise, a classification dataset for structural design problems will be severely

imbalanced, which is indicated from the values of β_T and P_f. As structural design problems often include an imbalanced dataset, it is challenging to select an adequate statistical metric that provides informative and truthful results.

This study employed three statistical measures to evaluate the performance including specificity, recall, and the Matthews correlation coefficients (MCC) [29]. For imbalanced class distributions, the majority class is typically referred to as the negative outcome and the minority class is assigned to the positive outcome. Therefore, structural failure is the positive outcome and no failure is the negative outcome. Specificity is the probability that an actual negative will test negative and is calculated by $TN/(TN+FP)$, which is the true negative rate. Specificity refers to how well a model identifies the frames which have no failure. Recall, also called the true positive rate, is the ratio of correct positive predictions to the total positive examples and is computed by $TP/(TP+FN)$. Recall informs how many positive predictions are missed from the prediction. The MCC is a reliable measure for imbalanced classification problems because it takes into account the ratio between positive and negative outcomes, which are not considered in both specificity and recall. The MCC is independent of the class imbalance, thus can reduce misleading results on imbalanced datasets [22]. The value of MCC varies between -1 and 1, similar to other correlation coefficients. The score is high only when all four categories in the confusion matrix are generated correctly. The MCC is computed by:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{10}$$

3 Reliability-based sensitivity analysis

Existing steel design specifications [33–36] provide guidance for inelastic analysis, also referred to as advanced analysis and GMNIA, which directly considers geometric and material nonlinearities and includes uncertainty in system, member, and connection strength and stiffness. Estimating structural performance with certainty is challenging because of the inherent uncertainty in a structural system which affects system performance. Reliability-based sensitivity analysis estimates the effect of an input variable by evaluating the structural performance with the variable under consideration as random while all other variables are at their nominal values. After repeated simulations for each property under consideration, the probability of failure P_f is estimated by n/N [37] where n = number of simulations which resulted in failure and N = total number of simulations. The system reliability index β is computed based on P_f. Unlike a reliability analysis which takes into account multiple random variables simultaneously, a sensitivity analysis considers only one random variable per simulation set to examine how the random variable affects the system behavior, therefore the strength distribution might have a smaller COV compared to that from a reliability analysis. The normal probability plot [38] is used to estimate P_f and β when no failure cases occur.

Researchers have investigated the system reliability of various steel structures estimated by considering uncertainties in the systems. Buonopane [39] conducted a reliability sensitivity study on two steel frames by considering uncertainties in yield strength, elastic modulus, residual stress, and sway and bow imperfections. Szyniszewski [40] investigated the effect of random geometric imperfections on progressive collapse propagation by analyzing 3-D steel framed buildings with uncorrelated geometric imperfections between structural members. Shayan et al. [41] presented a probabilistic study regarding modeling random geometric imperfections on regular and irregular sway and braced planar steel frames. Thai et al. [42] evaluated the system reliability of steel frames with semi-rigid connections. Uncertainties in gravity loads, material properties, cross-sectional properties, and connection properties were included in the reliability analysis. Zhang et al. [43] examined the system reliability of five steel structures including a beam, a portal frame, and three low-rise frames. Randomness considered in the analysis includes gravity loads, material properties, cross-sectional properties, and sway imperfection. Cardoso et al. [44] calibrated the system reliability of cold-formed steel portal frames with uncertain parameters in material properties, cross-section thickness, joint properties, and geometric imperfections.

4 Data collection

4.1 Structural system

Two example steel frames designed according to AISC 360 [34] are analyzed in this study, which have the same layout but different member sizes and loads, adopted from [45]. Fig. 1a and Table. 1 summarize the information of the example frames including geometry and applied loads. The frames are modeled in OpenSees [46] with displacement-based and fiber-type elements. All cross-sections contain the residual stress with the Galambos and Ketter pattern [47] and the fiber distribution as shown in Fig. 2. The nominal peak compressive residual stress value is $0.3F_{yn}$, where F_{yn} = nominal material yield strength of 248 MPa (36 ksi). Initial sway imperfection of $h/500$ was given to all columns in the frame, where *h* is the story height. The inelastic finite element (FE) analyses are conducted with all nominal properties to find the nominal value of the ultimate load ratio λ, which is the ratio of ultimate to factored design loads. The ultimate load ratios λ obtained from the analyses are 1.08 for both frames, i.e., they have a limited capacity for load redistribution. Fig. 1b and c illustrate the location of the most highly yielded zones for Frames 1 and 2, respectively. Frame 1 fails by the instability (inelastic buckling) of a ground floor column C2 and at failure members C2 and B2 are partially yielded. Frame 2 fails from a gradual sequence of yielding, thereby multiple members have highly yielded zones where a location along the member yielded greater than 75%, which are B1, B2, B3, B4, C5, and C6.
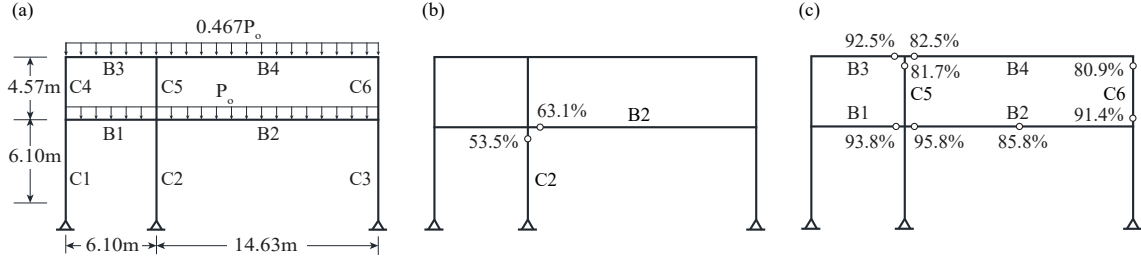

Figure 1: (a) Frame layout; location of highly yielded zones for (b) Frame 1; (c) Frame 2

Table 1: Member sizes and applied loads

Element	Frame 1	Frame 2
C1	W6 × 20	W12 × 14
C2	W14 × 82	W14 × 99
C3	W14 × 68	W14 × 82
C4	W6 × 8.5	W10 × 12
C5	W14 × 145	W14 × 109
C6	W14 × 145	W14 × 109
B1	W30 × 132	W27 × 84
B2	W36 × 182	W36 × 135
B3	W24 × 55	W18 × 40
B4	W30 × 116	W27 × 94
Loads (P_o)	111.86 kN/m	109.45 kN/m

4.2 Uncertainty

The Monte Carlo sampling method is used to generate samples of the uncertainties in material yield strength F_y, modulus of elasticity *E*, sway imperfection, and residual stress. Table 2 summarizes the statistical information and references of the uncertainties. Yield strength and elastic modulus are modeled following the distributions published in Bartlett et al. [48]. Nominal yield strength F_{yn} of 248 MPa and nominal elastic modulus E_n of 200 GPa are utilized to determine the mean value of yield strength and elastic modulus, respectively. The distribution of sway imperfection followed the distribution of Lindner and Gietzelt [49]. The scale factor of maximum compressive residual stress X is modeled as a normal distribution provided in Shayan et al. [50]. The random scale factor X was multiplied by σ_{RC} to consider the uncertainty of residual stress magnitudes. The peak tensile residual stress (σ_{RT}) within a cross section was determined by the geometry and the peak compressive residual stress (σ_{RC}), which equals $0.3F_{yn}$, as shown in (Fig. 2). σ_{RT} includes X indirectly because it is calculated based on equilibrium. Once σ_{RC} and σ_{RT} are determined,
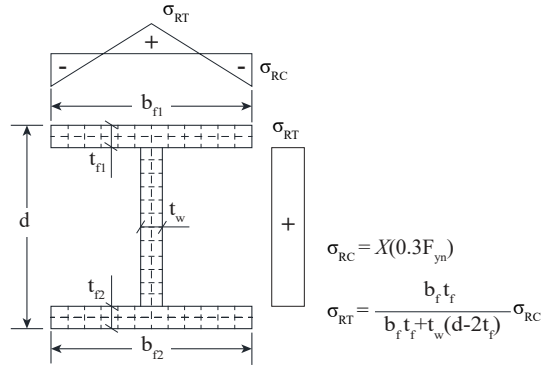
Figure 2: Residual stress pattern and fiber distribution

the rest of the residual stresses in the cross section were set based on the residual stress pattern. The residual stress condition is constant along the length of a member. To account for the maximum number of uncertainties and investigate the effect of each parameter on structural failure, the frames are assumed as spatially uncorrelated, i.e., all structural members have different random properties.

Table 2: Description of feature variables

Variable	Mean	COV	Distribution	References
F_y	$1.1F_{yn}$	0.06	Lognormal	Bartlett et al. [48]
E	E_n	0.04	Lognormal	Bartlett et al. [48]
Sway imperfection	1/770	0.875	Lognormal	Lindner and Gietzelt [49]
X	1.064	0.27	Normal	Shayan et al. [50]

4.3 Dataset

As the structural members are uncorrelated, each individual member had a different realization of the random properties. In other words, there were no identical random values of the input variables shared between all beams or all columns. For example, input feature variables consist of thirty-three different parameters including ten different values of yield strength, elastic modulus, and residual stress and three sway imperfections assigned to the three column locations – left, center, and right. The response variable is the binary outputs based on the ultimate load ratio λ obtained from the FE analysis containing random realizations of the input parameters. As the frame is designed according to the inelastic method provided in AISC 360 [34], this study applied the probability-based limit state design criteria $\lambda = 1.0$ as the classification criteria of the dataset. If the ultimate load ratio is less than 1.0, the frame experiences a structural failure and the observation is assigned to Class 1. An ultimate load ratio greater than or equal to 1.0 indicates that the frame is safe and the sample set is assigned to Class 0, which means no failure. A total of 500,000 simulations were run for Frame 1 and 1,000,000 simulations for Frame 2. A small percent of simulations had convergence issues and were excluded from the datasets. In total, Frame 1 had 309 failures (Class 1) out of 498,050 labeled data points. Frame 2 had 127 failures, which is less than that of Frame 1, out of the total number of simulations, 903,272. The data points for each frame were randomly assigned into equal training and test sets, for a 50%-50% split between training and testing datasets for each frame.

As discussed previously, a classification dataset for structural design problems will be severely imbalanced, based on the selection of β_T and the corresponding P_f. The no-information rate describes how much the dataset is imbalanced, which is calculated by $max(n_{Class\,0}, n_{Class\,1})/(n_{Class\,0} + n_{Class\,1})$, where $n_{Class\,0}$ is the number of Class 0 examples and $n_{Class\,1}$ is the number of Class 1 examples. The no-information rates of Frame 1 and Frame 2 are 99.94% and 99.99%, respectively, approximately equal to 100%. As imbalanced classification data leads to biased prediction toward the majority class, this study employed the oversampling method for the training set to reduce the problem caused by imbalanced classification. The minority class is oversampled to have the same number as the majority class by the Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al. [51]. For example, Frame 2 had 7,111 times greater number of the majority class than the minority class. The new minority class examples are generated 7,111 times in a training set to equal the number of majority class examples. The oversampled

minority class are not duplicates of the existing samples, but are derived using k-nearest neighbors and interpolation parameters.

5   Comparison of sensitivity analysis results

This section compares the sensitivity analysis results obtained by the feature importance approach and the reliability sensitivity analysis. For the reliability-based sensitivity study, 40,000 simulations for each uncertainty under consideration were conducted. The feature name consists of the structural member name following the property name. Residual stress and sway imperfection are shortened to 'rs' and 'sway', respectively. For example, E-B2 represents the elastic modulus of B2 and rs-C1 indicates the residual stress of C1. Sway imperfections at the column locations – left, center, and right – are labeled as sway-C1, sway-C2, and sway-C3, respectively.

Table 3 summarizes the reliability-based sensitivity analysis results including statistics of strength, probability of failure, and reliability index. Frame 1 has larger COVs than Frame 2, with a larger difference for random elastic modulus and sway imperfection, which indicates that Frame 1 is more sensitive to these factors. Although the nominal ultimate strength was equal to 1.08 for both frames, Frame 1 has smaller values of β compared to those of Frame 2 due to the larger COVs of Frame 1 resulting in a lower boundary of the strength distributions.

Table 3: Reliability-based sensitivity results for Frames 1 and 2

| | Frame 1 | | | | Frame 2 | | | |
| | Strength | | Reliability | | Strength | | Reliability | |
Random variable	Mean	COV	β	P_f	Mean	COV	β	P_f
Yield strength	1.13	0.034	3.41	3.3×10^{-4}	1.16	0.031	3.79	7.5×10^{-5}
Elastic modulus	1.07	0.007	9.01	≈ 0	1.07	0.004	16.8	≈ 0
Sway imperfection	1.01	0.030	2.65	4.0×10^{-3}	1.08	0.011	3.51	2.3×10^{-4}
Residual stress	1.07	0.003	27.7	≈ 0	1.07	0.002	37.2	≈ 0

5.1   Frame 1: instability of a single column

The results of the feature importance method are shown in Fig. 3. The top row (Fig. 3a-c) shows the top ten feature rankings of Frame 1 derived from the data analysis methods including ANOVA, mRMR, and Spearman's rank. The feature orders obtained from the model analysis techniques including impurity-based, permutation, and SHAP are shown in the bottom row (Fig. 3d-f). As previously discussed in Section 2, the feature importance techniques can derive either positive or negative values or both. As the feature rankings show only ten highly-ranked features, the negative scores are not included in the figure except for Spearman's rank (Fig. 3c), which rated the features based on their feature importance value magnitude. Sway-C2 and F_y-C2 are top-ranked from all the feature importance methods. Sway-C3 or E-C3 is third-ranked but has a negligible importance score in comparison to the top two features. Although the order of the remainder of the features is different between the various techniques, only the two highly-ranked features have significant scores. In other words, only the first two features are significant to the prediction of failure for Frame 1. Most methods result in scores approximately equal to zero for the least important features. Frame 1 fails by the inelastic instability of C2, and this is reflected in the importance score results as the features related to C2 are the most highly ranked.

From the results of Table 3, random sway imperfection resulted in the lowest β in Frame 1, followed by yield strength, elastic modulus, and residual stress. A small elastic modulus and large sway imperfection increase lateral deflections, thereby increasing second-order bending moments. As Frame 1 fails by the instability of C2, the frame capacity is most influenced by the factors resulting in increased bending moments.

Fig. 4 illustrates the scatter plots of input random properties versus the frame strength based on the reliability-based sensitivity studies. As shown in Fig. 4a, the yield strength of C2 and the frame strength have a nearly perfect correlation, which indicates that the strength of Frame 1 is controlled by C2. The yield strength of the other members showed no correlation with frame strength. Fig. 4b and c show that the frame strength has a weak correlation with the elastic modulus of C3 and the residual stress of C2. The sway imperfection at the center column (C2) has the most significant impact on the strength among the three column positions — left (C1), center, and right (C3) — by showing a strong positive correlation.
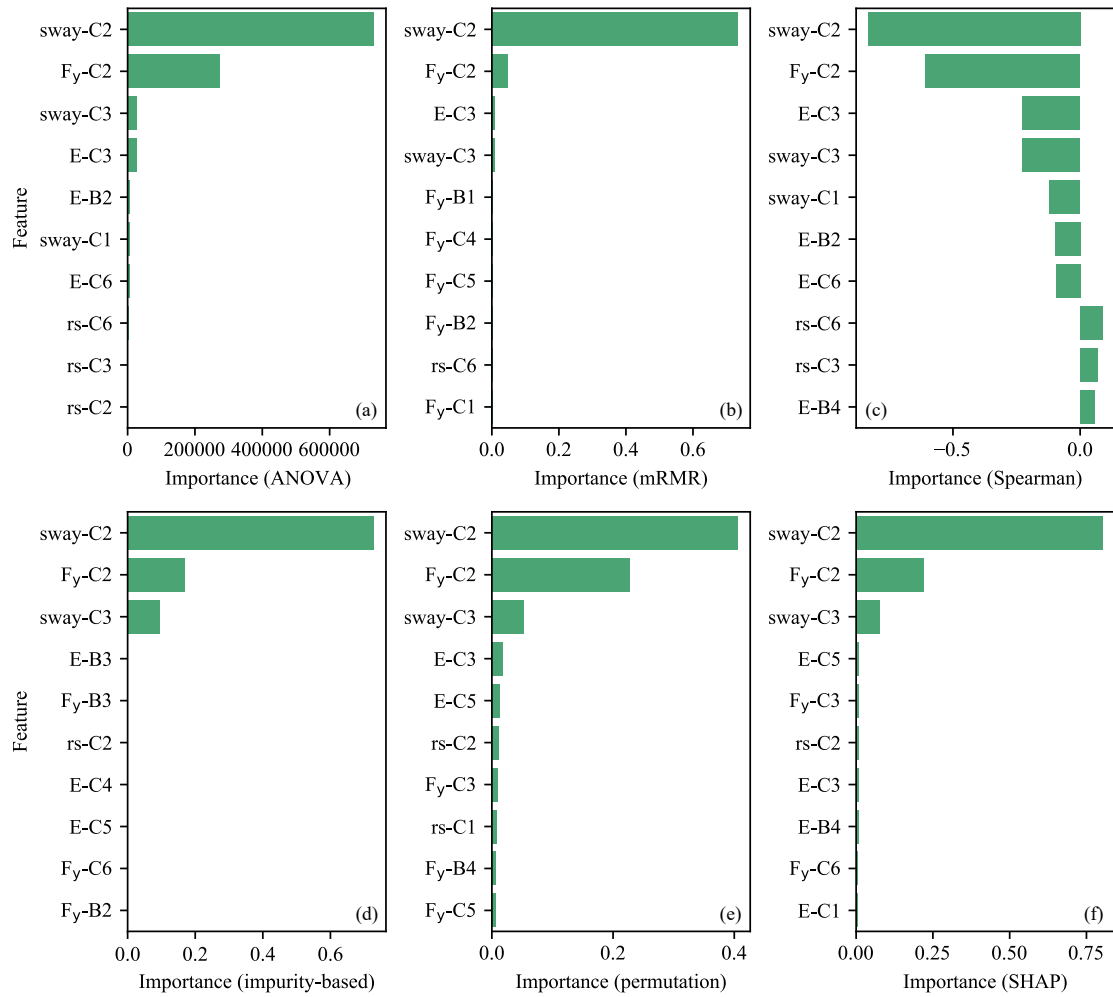
Figure 3: Importance ranking of the top ten features of Frame 1 derived by data analysis methods (top row) and model analysis methods (bottom row)

The highly-ranked features determined from the feature importance methods are identical with the factors that resulted in a lower β from the reliability-based sensitivity analysis. The feature rankings determined by the feature importance framework showed that sway-C2 and F_y-C2 are the most important features among the thirty-three random properties. The third-ranked feature is either sway-C3 or E-C3, which showed a positive correlation with the frame strength, but less significant than sway-C2 and F_y-C2. Overall, the random properties that have significant impacts on Frame 1's capacity determined by the reliability-based and machine learning-based sensitivity analyses are in agreement.

5.2 Frame 2: progressive yielding

Fig. 5 shows the top ten feature rankings of Frame 2 derived by the feature importance approach. As shown in Fig. 5a-c, the top-ranked feature is either F_y-C6 or sway-C2, and the remaining order varies for each data analysis technique, which identifies important features without model fitting. However, the model analysis techniques, which require model training to measure feature importance, derived the same top four features including sway-C2, F_y-C6, F_y-B2, and sway-C3 in descending order (Fig. 5d-f). The features ranked fourth through tenth are similar between the model analysis methods. At least four yield strengths are highly-ranked across all the methods, which indicates that yield strength is an influential factor in the failure of Frame 2 and the failure mode is progressive yielding. As previously shown in Fig. 1c, Frame 2 has six members that have critical impacts on the system failure, including four beams and two columns. In particular, B2 and C6 have two highly yielded zones each, and the yield strengths of these members are top-ranked among all the yield strengths. The feature importance results show the significant members in the
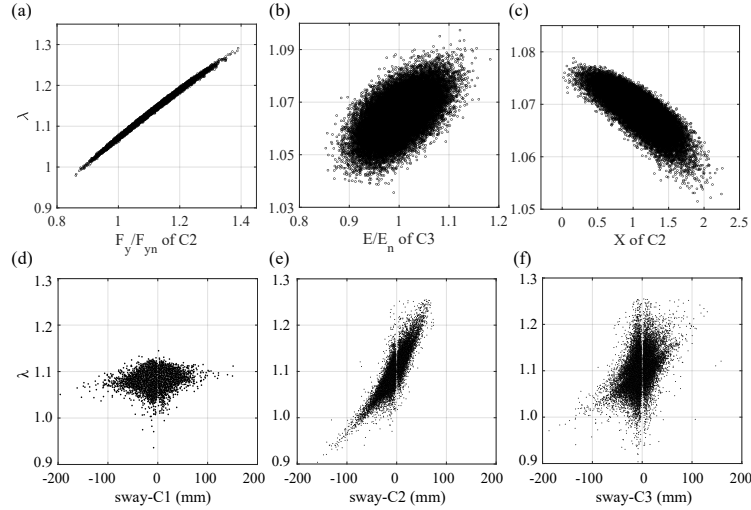
Figure 4: Scatter plots of Frame 1 strength versus random (a) F_y of C2 (b) E of C3 (c) X of C2 (d) sway imperfection of C1 (e) C2 (f) C3

system in addition to the influential properties. Due to the complex failure mode of Frame 2, the feature orders are not as straightforward as Frame 1, however the results indicate that the feature importance approach is accurate for steel frames with various failure modes.

The reliability-based sensitivity study investigated the effects of random properties on the frame strength. As shown in Fig. 6, the properties that have a significant effect are identical with the features that are highly ranked by the feature importance methods. Fig. 6a and b show dents on the upper left side, and they occur when the value of the B2 yield strength is the maximum or the minimum among all members, respectively. Fig. 6c and d indicate that Frame 2's strength has positive correlations with elastic moduli of C2 and B2. As X of C2 increases, the frame strength decreases (Fig. 6e) because the presence of residual stresses leads to the onset of yielding at a lower applied load [39]. Random elastic modulus and residual stress of C2 and B2 are correlated with the frame strength but showed small COVs, which represents less significant influence. The effects of random sway imperfection are shown in Fig. 6f-h. The center columns sway have the most significant impact on the frame strength than the sway of other columns. This example illustrates that not all factors influence the system behavior, and it is therefore unnecessary to assess the effects of each factor individually, as is done in a reliability-based sensitivity analysis. On the other hand, the feature importance approach analyzes all the factors at once to estimate the effects on system behavior.

Overall, Frame 2 has smaller magnitudes of the importance scores than for Frame 1. Moreover, Frame 1 has a large difference between the two top-ranked features and the remainder of the features, while the score difference between the features in Frame 2 is smaller. In other words, the importance score of Frame 2 decreases smoothly from the top to the bottom of the rankings. When a structural system fails by a single member (Frame 1), the properties of that member has a critical impact on the entire system. On the other hand, when various members lead to a system failure, such as progressive yielding (Frame 2), the properties of multiple members have a significant impact on the entire system. A comparison of the importance values between Frame 1 and Frame 2 indicates that the number of structural members involved in system failure influences the magnitude of importance as well as the number of features considered to be important.

6   Performance evaluation

The test set, which accounts for 50% of the dataset, is employed to evaluate the models fitted on the training set. The accuracy metrics include specificity and recall to measure the correct prediction of Class 0 (no failure) and Class 1 (failure), respectively. In addition, the Matthews correlation coefficient (MCC) is used, which is a suitable metric for the imbalanced dataset.

The predictive performance of the machine learning models for Frame 1 is shown in Fig. 7. The specificity curves for both logistic regression and decision tree algorithms (Fig. 7a and b) show that the specificity
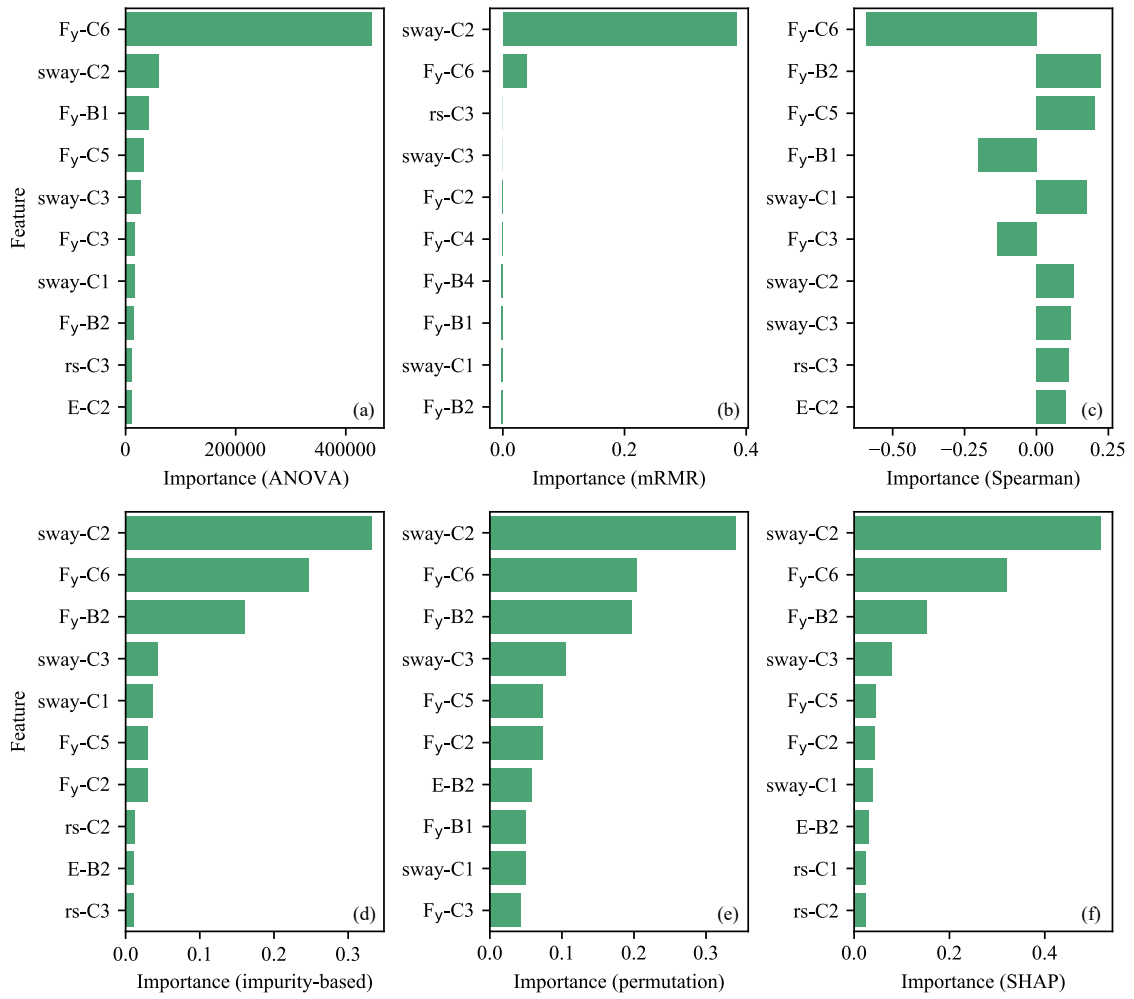
Figure 5: Importance ranking of the top 10 features of Frame 2 derived by data analysis methods (top row) and model analysis methods (bottom row)
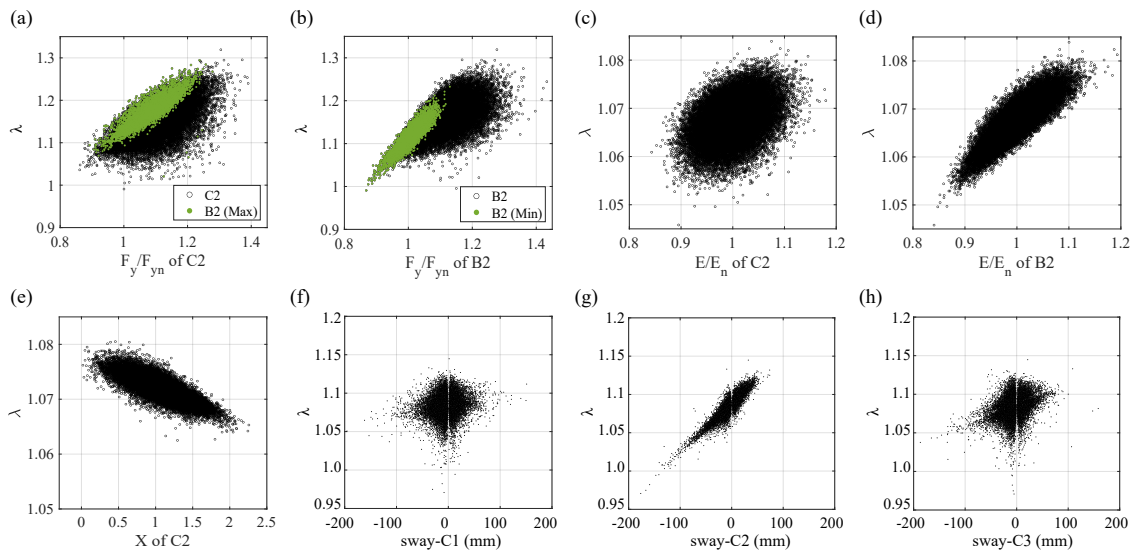

Figure 6: Scatter plots of Frame 2 strength versus random (a) F_y of C2 (b) F_y of B2 (c) E of C2 (d) E of B2 (e) X of C2 (f) sway imperfection of C1 (g) C2 (h) C3

score reaches the nearly perfect value of approximately 1.0, even with only a few features. In this study, specificity represents the proportion of safe structures that are correctly predicted. The Frame 1 dataset is severely imbalanced with the no-information rate of 99.94%, therefore the model performance measured by specificity shows inflated results due to the biased classification towards the majority class. Recall is computed to evaluate how good a model is at detecting a structural failure, which is the positive class. As it is critical to identify system failure rather than safe structures in structural design practice, recall is a more crucial measure than specificity in this study. Fig. 7c shows the recall curve obtained from the logistic regression model. When the feature set contains the top three features of the model analysis techniques and ANOVA, which are sway-C2, F_y-C2, and sway-C3, the recall score rapidly increases up to 0.77. Spearman and mRMR ranked the sway imperfection of C3 at fourth and therefore the score abruptly rises when the feature set increases to four. The recall curve of the decision tree (Fig. 7d) shows the highest value when only two or three features are selected. The outcome recall scores converge to a lower score of 0.72 after reaching the peak point. Fig. 7e and f show the outcome MCC scores for the logistic regression and decision tree models, respectively. The MCC curves have a similar shape as the recall curves; the logistic regression model performance improves as the feature set increases, and the decision tree model reaches the peak point when the feature set is small. This indicates an overfitting issue which occurs when the model is trained on a large feature set. Feature selection based on the feature importance score can improve the overfitting by excluding the redundant features from training. The least important features, which are ranked after the fifteenth, could be removed to obtain a better performance.

The model performance of Frame 2 measured by specificity is shown in Fig. 8a and b. The specificity generates overoptimistic results due to the high imbalanced ratio of 99.99%. When the yield strength of C6 is ranked as the most important feature by ANOVA or Spearman, the decision tree model shows the nearly-perfect score even though the dataset includes only one feature (Fig. 8b). The lowest score of specificity is 0.92, indicating that both logistic regression and decision tree models can identify safe structures with high accuracy. The recall curves of Frame 2 are shown in Fig. 8c and d. The logistic regression model can correctly predict the failure only when the sway imperfection of C2 is included in the feature set. For example, the recall curve of Spearman's rank shows zeros until the feature set size is seven because the feature ranking rated the sway imperfection of C2 at seventh. The recall curve of the decision tree model (Fig. 8d) shows a large variation between the data analysis methods because they had completely different feature rankings. However, each curve merges to about 0.32 as the number of features increases. The MCC curve of the logistic regression model (Fig. 8e) shows a similar shape as the recall curve because the specificity scores had a single value, which is close to 1, regardless of the feature set size. The MCC values of the decision tree model merge to 0.15, which is the score of the entire feature set (Fig. 8f). When a high-dimensional dataset is used, the class imbalance leads to additional challenges in misclassification of the minority class [21]. As previously discussed in Section 4.3, Frame 2 consisted of a larger sample space and a fewer number of failures than for Frame 1. Moreover, the feature orders of Frame 2 were inconsistent between the feature importance techniques, as multiple features were significant to the prediction of system failure. The higher imbalanced ratio and the complex failure mode of Frame 2 result in a low performance measured by recall as well as the MCC. The extremely imbalanced classification of Frame 2 led to a lower performance for predicting structural failure because the machine learning classifiers had only a few minority class examples to oversample in the training set as well as to test the model prediction. The model performance, which measures the prediction of minority class, can be improved by obtaining more number of failures in a dataset or reducing the class imbalanced ratio.

The six feature importance techniques showed similar performance of Frame 1 measured by the MCC. The performance was improved after containing the three top-ranked features (Fig. 7e and f). In Frame 2, however, the model analysis methods showed more accurate results than the data analysis methods. In particular, the permutation and SHAP methods showed the best performance until the feature set size increases to seven (Fig. 8f) because they ranked F_y-C5 and F_y-C2 at fifth and sixth, respectively, whereas the impurity-based method ranked F_y-C2 at seventh, and F_y-C5 and F_y-C2 have a significant influence on the system failure of Frame 2. In summary, based on the feature rankings and the model performance results of both frames, SHAP and permutation methods are the best techniques for estimating the importance of features.

7   Conclusion

This study examined the feature importance approach using datasets with a large number of uncertainties and severely imbalanced classification. Two designs of a non-symmetric planar steel frame were investi-
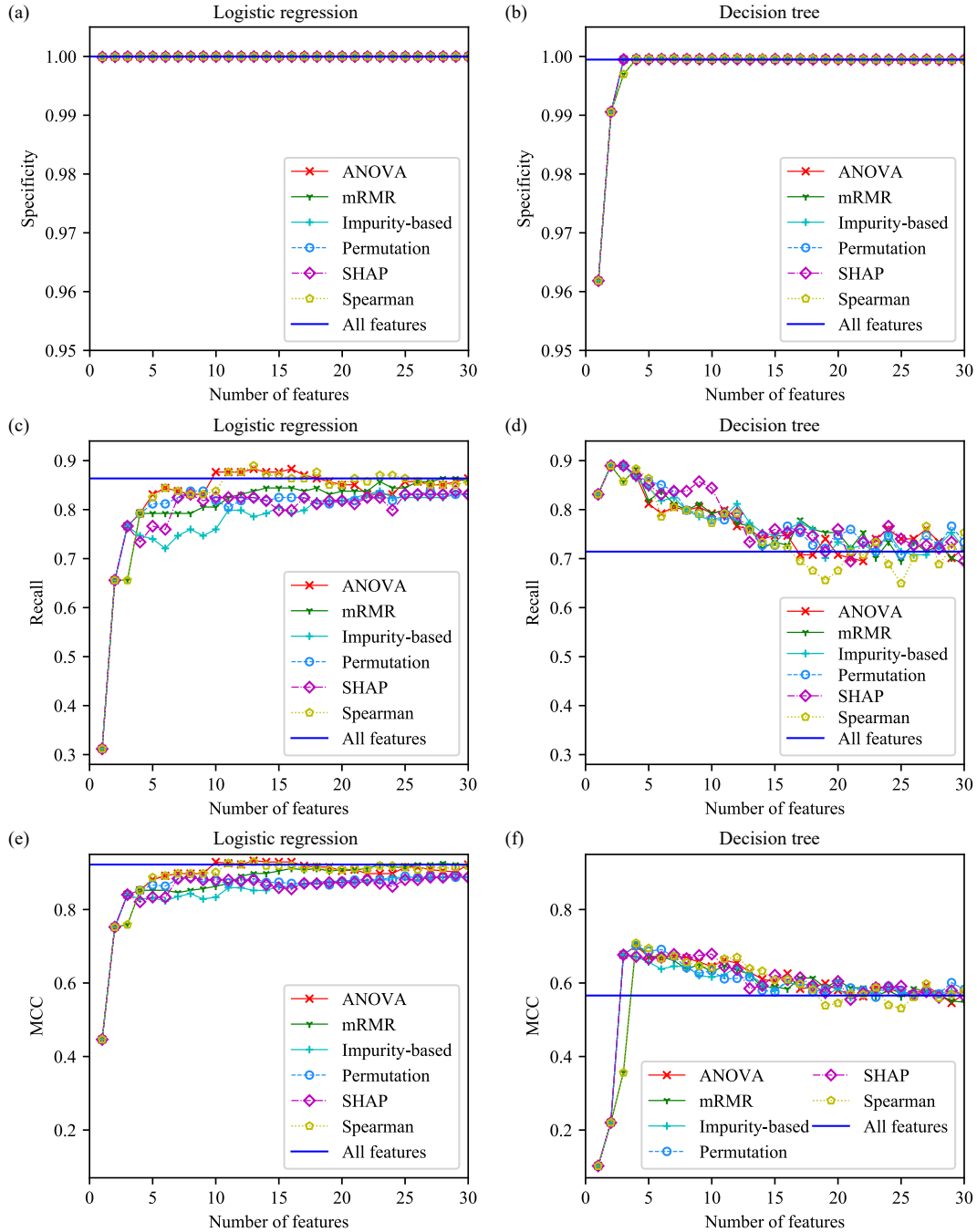
Figure 7: Frame 1 specificity, recall, and MCC for logistic regression (left column) and decision tree (right column)

gated with consideration of uncertainties in material yield strength, Young's modulus, sway imperfection, and residual stress. The dataset information consisted of thirty-three uncertainties based on the uncorrelated scenario and the ultimate load ratio obtained from the finite element analysis. A scarce number of failures occurred, as is common in structural engineering design, thus the datasets were extremely class-imbalanced with the two classes being safe and fail. Feature importance techniques including ANOVA, mRMR, Spearman's rank, impurity-based, permutation, and SHAP are trained on the high-dimensional and severely class-imbalanced datasets to identify the important features. After rating the features according to the importance score, the logistic regression and decision tree algorithms are trained to predict the classes using the feature set containing the top-ranked features. The model performance evaluated by specificity showed the nearly-perfect performance for both frames because most examples were assigned to the majority class, safe structure. Frame 1, which failed by the buckling of a single column, showed
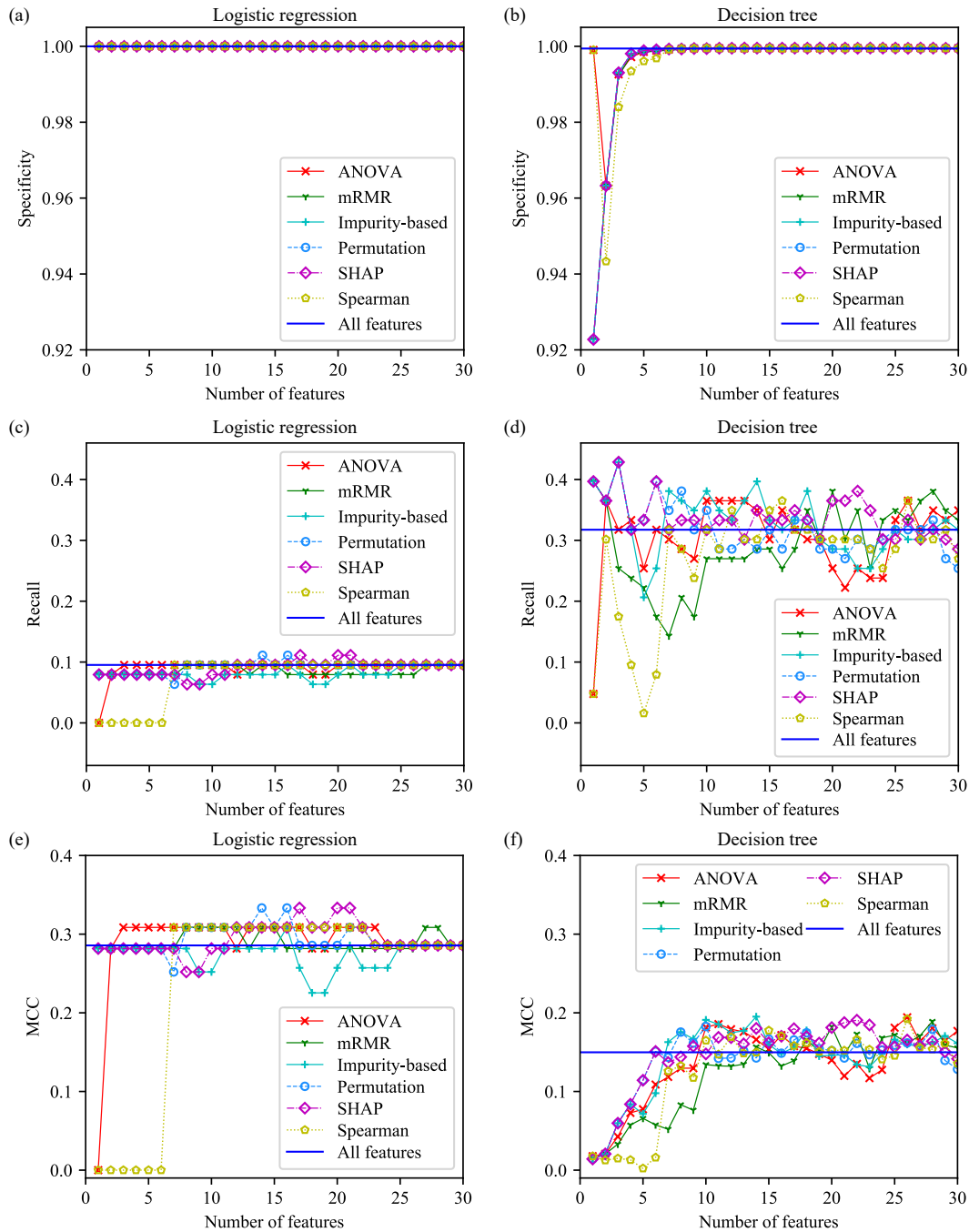
Figure 8: Frame 2 specificity, recall, and MCC for logistic regression (left column) and decision tree (right column)

good performance for the prediction of failure even with the highly imbalanced classification. However, for Frame 2 which had a complex failure mode of progressive yielding in addition to an extremely low failure probability, it was challenging to obtain the accurate prediction of the minority class. As class-imbalanced data is inevitable in structural engineering, it is necessary to be cautious in assessing the predictive accuracy of structures with a complex failure mode and a failure probability approximately equal to zero. The important features identified using the machine learning based feature importance approach were compared with the results of a conventional reliability-based sensitivity study to identify factors which result in a lower system reliability index. Overall, both methods identified the same factors which reflected the system failure modes. This study demonstrated that the machine learning-based sensitivity analysis can identify the influential features affecting system failure even when there are high-dimensional uncertain parameters and a highly imbalanced dataset.

References

[1] Y.-T. Wu, "Computational methods for efficient structural reliability and reliability sensitivity analysis," *AIAA J*, vol. 32, no. 8, pp. 1717–1723, 1994.

[2] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*, vol. 10. John Wiley & Sons, 2016.

[3] A. J. Torii, R. H. Lopez, and L. F. F. Miguel, "Probability of failure sensitivity analysis using polynomial expansion," *Probabilistic Eng Mech*, vol. 48, pp. 76–84, 2017.

[4] C. Proppe, "Local reliability based sensitivity analysis with the moving particles method," *Reliab Eng Syst Saf*, vol. 207, p. 107269, 2021.

[5] M. R. Sheidaii and R. Bahraminejad, "Evaluation of compression member buckling and post-buckling behavior using artificial neural network," *J Constr Steel Res*, vol. 70, pp. 71–77, 2012.

[6] Y. Pu and E. Mesbahi, "Application of artificial neural networks to evaluation of ultimate strength of steel panels," *Eng Struct*, vol. 28, no. 8, pp. 1190–1196, 2006.

[7] S. Chiew, A. Gupta, and N. Wu, "Neural network-based estimation of stress concentration factors for steel multiplanar tubular xt-joints," *J Constr Steel Res*, vol. 57, no. 2, pp. 97–112, 2001.

[8] M. Dabiri, M. Ghafouri, H. R. Raftar, and T. Björk, "Utilizing artificial neural networks for stress concentration factor calculation in butt welds," *J Constr Steel Res*, vol. 138, pp. 488–498, 2017.

[9] V. V. Degtyarev, "Neural networks for predicting shear strength of cfs channels with slotted webs," *J Constr Steel Res*, vol. 177, p. 106443, 2021.

[10] J. Rahman, K. S. Ahmed, N. I. Khan, K. Islam, and S. Mangalathu, "Data-driven shear strength prediction of steel fiber reinforced concrete beams using machine learning approach," *Eng Struct*, vol. 233, p. 111743, 2021.

[11] D.-C. Feng, W.-J. Wang, S. Mangalathu, G. Hu, and T. Wu, "Implementing ensemble learning methods to predict the shear strength of rc deep beams with/without web reinforcements," *Eng Struct*, vol. 235, p. 111979, 2021.

[12] M. Almustafa and M. Nehdi, "Machine learning model for predicting structural response of rc slabs exposed to blast loading," *Eng Struct*, vol. 221, p. 111109, 2020.

[13] J. Kiani, C. Camp, and S. Pezeshk, "On the application of machine learning techniques to derive seismic fragility curves," *Comput Struct*, vol. 218, pp. 108–122, 2019.

[14] S.-H. Hwang, S. Mangalathu, J. Shin, and J.-S. Jeon, "Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames," *J Build Eng*, vol. 34, p. 101905, 2021.

[15] S. Mangalathu, H. Jang, S.-H. Hwang, and J.-S. Jeon, "Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls," *Eng Struct*, vol. 208, p. 110331, 2020.

[16] H. Huang and H. V. Burton, "Classification of in-plane failure modes for reinforced concrete frames with infills using machine learning," *J Build Eng*, vol. 25, p. 100767, 2019.

[17] H. Koh and H. B. Blum, "Machine learning-based feature importance approach for sensitivity analysis of steel frames," *engrXiv*, 2021.

[18] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *J Mach Learn Res*, vol. 15, no. 1, pp. 3133–3181, 2014.

[19] P. Orzechowski, W. La Cava, and J. H. Moore, "Where are we now? a large benchmark study of recent symbolic regression methods," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1183–1190, 2018.

[20] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural comput*, vol. 8, no. 7, pp. 1341–1390, 1996.

[21] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC bioinform*, vol. 11, no. 1, pp. 1–17, 2010.

[22] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genom*, vol. 21, no. 1, pp. 1–13, 2020.

[23] T. Parr, J. D. Wilson, and J. Hamrick, "Nonparametric feature impact and importance," *arXiv preprint arXiv:2006.04750*, 2020.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[25] C. Spearman, "The proof and measurement of association between two things," *Am J Psychol*, vol. 15, no. 1, pp. 72–101, 1904.

[26] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[27] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[28] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[29] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochim Biophys Acta*, *Protein Struct*, vol. 405, no. 2, pp. 442–451, 1975.

[30] A. J. Izenman, *Modern multivariate statistical techniques: Regression, classification and manifold learning*. New York City, NY: Springer, 2013.

[31] B. Ellingwood, "Probability-based codified design: past accomplishments and future challenges," *J Struct Saf*, vol. 13, no. 3, pp. 159–176, 1994.

[32] ASCE, Reston, VA, *Minimum design loads for buildings and other structures*, 2016.

[33] EN 1993-1-1, *Eurocode 3: Design of steel structures - Part 1-1: General rules and rules for buildings*. European Committee for Standardisation, 2005.

[34] ANSI/AISC, Chicago, IL, *Specification for Structural Steel Buildings*, 2016.

[35] AS 4100, *Steel Structures*. Standards Australia, 1998.

[36] CSA S16:19, *Design of Steel Structures*. Canadian Standards Association, 2019.

[37] R. Melchers and A. T. Beck, *Structural reliability analysis and prediction*. Hoboken, NJ: John Wiley & Sons Ltd, 3rd ed., 2018.

[38] J. M. Chambers, *Graphical methods for data analysis*. CRC Press, 1983.

[39] S. G. Buonopane, "Strength and reliability of steel frames with random properties," *J Struct Eng*, vol. 134, no. 2, pp. 337–344, 2008.

[40] S. Szyniszewski, "Effects of random imperfections on progressive collapse propagation," in *Structures Congress 2010*, pp. 3572–3577, 2010.

[41] S. Shayan, K. J. Rasmussen, and H. Zhang, "On the modelling of initial geometric imperfections of steel frames in advanced analysis," *J Constr Steel Res*, vol. 98, pp. 167–177, 2014.

[42] H.-T. Thai, B. Uy, W.-H. Kang, and S. Hicks, "System reliability evaluation of steel frames with semi-rigid connections," *J Constr Steel Res*, vol. 121, pp. 29–39, 2016.

[43] H. Zhang, H. Liu, B. R. Ellingwood, and K. J. Rasmussen, "System reliabilities of planar gravity steel frames designed by the inelastic method in aisc 360-10," *J Struct Eng*, vol. 144, no. 3, p. 04018011, 2018.

[44] F. S. Cardoso, H. Zhang, K. J. Rasmussen, and S. Yan, "Reliability calibrations for the design of cold-formed steel portal frames by advanced analysis," *Eng Struct*, vol. 182, pp. 164–171, 2019.

[45] R. D. Ziemian, *Advanced methods of inelastic analysis in the limit states design of steel structures*. PhD thesis, Cornell University, Ithaca, NY, 1990.

[46] S. Mazzoni, F. McKenna, M. H. Scott, G. L. Fenves, *et al.*, *OpenSees command language manual*. University of California, Berkeley, 2006.

[47] T. V. Galambos and R. L. Ketter, "Columns under combined bending and thrust," *J Eng Mech Div*, vol. 85, no. 2, pp. 1–30, 1959.

[48] F. M. Bartlett, R. J. Dexter, M. D. Graeser, J. J. Jelinek, B. J. Schmidt, and T. V. Galambos, "Updating standard shape material properties database for design and reliability," *Eng J AISC*, vol. 40, no. 1, pp. 2–14, 2003.

[49] J. Lindner and R. Gietzelt, "Imperfektionsannahmen für stützenschiefstellungen," *Der Stahlbau*, vol. 52, no. 4, pp. 97–101, 1984.

[50] S. Shayan, K. J. R. Rasmussen, and H. Zhang, "Probabilistic modelling of residual stress in advanced analysis of steel structures," *J Constr Steel Res*, vol. 101, pp. 407–414, 2014.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J Artif Intell Res*, vol. 16, pp. 321–357, 2002.