



Research article

Enhancing structural health monitoring with machine learning for accurate prediction of retrofitting effects

A. Presno Vélez, M. Z. Fernández Muñiz* and J. L. Fernández Martínez

Department of Mathematics, Group of inverse problems, optimization and machine learning, Oviedo University, c/Federico García Lorca 18, Oviedo 33007, Spain

* **Correspondence:** Email: zulima@uniovi.es.

Abstract: Structural health in civil engineering involved maintaining a structure's integrity and performance over time, resisting loads and environmental effects. Ensuring long-term functionality was vital to prevent accidents, economic losses, and service interruptions. Structural health monitoring (SHM) systems used sensors to detect damage indicators such as vibrations and cracks, which were crucial for predicting service life and planning maintenance. Machine learning (ML) enhanced SHM by analyzing sensor data to identify damage patterns often missed by human analysts. ML models captured complex relationships in data, leading to accurate predictions and early issue detection. This research aimed to develop a methodology for training an artificial intelligence (AI) system to predict the effects of retrofitting on civil structures, using data from the KW51 bridge (Leuven). Dimensionality reduction with the Welch transform identified the first seven modal frequencies as key predictors. Unsupervised principal component analysis (PCA) projections and a K-means algorithm achieved 70% accuracy in differentiating data before and after retrofitting. A random forest algorithm achieved 99.19% median accuracy with a nearly perfect receiver operating characteristic (ROC) curve. The final model, tested on the entire dataset, achieved 99.77% accuracy, demonstrating its effectiveness in predicting retrofitting effects for other civil structures.

Keywords: health monitoring; retrofitting; machine learning

Mathematics Subject Classification: 68T45, 93C95

1. Introduction

Structural health, in the context of civil engineering and construction, refers to the ability of structures to maintain their integrity and performance over time, resisting loads and environmental effects without suffering significant damage. Deteriorated or defective infrastructure can lead to severe accidents, loss of life and property, interruptions in essential services, and negative economic impacts.

Therefore, it is essential to maintain and improve the condition of existing structures for their long-term functionality and sustainability. For this reason, structural health monitoring (SHM) is fundamental and is a common practice nowadays [1, 2].

Machine learning (ML) techniques have demonstrated significant promise in enhancing damage prediction within SHM systems. These techniques efficiently sift through vast amounts of sensor data gathered from SHM systems to identify relevant features and patterns indicative of damage, often escaping immediate detection by human analysts. ML models frequently outperform traditional methodologies, especially in deciphering intricate and nonlinear relationships between sensor data and structural health, leading to more accurate predictions [3, 4]. By leveraging historical data, ML algorithms can detect subtle shifts in structural behavior that might precede visible damage, enabling early detection of potential issues and paving the way for proactive maintenance to avert catastrophic failures [5]. Moreover, ML algorithms evolve and refine their predictive capabilities over time by assimilating new data, thereby strengthening their performance as they gain a deeper understanding of the structure's behavior. These techniques integrate seamlessly with sensor networks deployed for SHM, streamlining data processing and decision-making without manual intervention. ML-driven SHM systems support predictive maintenance strategies, reducing downtime and repair costs by flagging and addressing potential damage before it escalates [6, 7]. However, it is important to note that ML models are highly dependent on the quality of the data used for training.

Research in the field of structural health covers a broad range of topics, from techniques for monitoring and diagnosing the health of structures to methods of repair and rehabilitation. SHM is conducted using sensors and remote monitoring systems that detect vibrations, deformations, corrosion, cracks, and other indicators of deterioration, allowing real-time assessment of the structure's condition and characterization of defects and damage. The goal of SHM is to predict the service life of the structure and plan effective maintenance and rehabilitation strategies [8].

Given the growing importance of SHM in recent decades, as evidenced in related literature [9–11], a notable challenge is the scarcity of labeled data on damage states. The primary aim is to build a predictive model that learns from the available data, with the capability to forecast whether the retrofitting process has been effective. Retrofitting, a common practice in this field, involves strengthening, upgrading, or modifying existing structures to enhance their performance and durability. This technique is frequently employed to increase the resilience of buildings and infrastructure against risks such as earthquakes, floods, or deterioration caused by time [12, 13].

Initially, the model should be able to detect deviations in the structure's behavior from its normal operating mode, which can be achieved using data from the undamaged structure (an unsupervised anomaly detection problem). A drawback of this approach is that the dynamics of a structure's behavior can vary due to normal operational loads or environmental conditions without indicating damage. To eliminate the influence of these benign changes, data normalization is required [9, 14]. If the model is required to diagnose the progression and location of damage, supervised learning is necessary, thus requiring labeled data on damage states to identify them [9]. To address this data deficit, some authors [15, 16] propose knowledge transfer between similar structures within a homogeneous population and even between heterogeneous populations through transfer learning strategies.

Large structures built by humans are meant to last many years due to their economic cost, but damage caused by time and service often necessitates retrofitting. In the context of civil engineering and construction, retrofitting involves strengthening a structure to improve its resistance, upgrading

obsolete or inefficient systems, or adapting the structure for new uses to enhance or restore its safety, efficiency, or functionality. This procedure is typically associated with changes in the modal parameters of a structure, such as natural frequencies, which result from modifications to its stiffness [17].

Long-term damage detection can be performed using unsupervised classifiers, as it is impractical to cause real damage to structures due to their high cost and the danger of operating with a damaged structure. Therefore, given the lack of properly labeled training and testing data, transfer learning can be used to leverage knowledge gained from data collected before retrofitting and apply it to another related situation (data collected after retrofitting). This allows previously acquired knowledge to be utilized in the long-term damage detection process, even when significant changes occur in the structure due to retrofitting [13].

The aim of this work is to train an algorithm capable of predicting structural damage, for which data collected before, during, and after retrofitting of the KW51 bridge in Leuven, Belgium will be used [17]. To predict the effect of retrofitting, we first applied dimensionality reduction by transforming the accelerograms to the spectral domain using the Welch transform [18]. The Fisher ratio of the decimal logarithm of the power spectral density (PSD) for each of the bridge's modal frequencies was found to have discriminative power, identifying the frequencies that best predict the retrofitting effect. The first seven modal frequencies were the most discriminative. These results were confirmed using a sample of 100 evenly distributed frequencies within the range of the modal frequencies.

The unsupervised principal component analysis (PCA) projection in 2D and 3D spaces, formed by the first three principal components obtained from the decimal logarithm of the power spectral densities calculated at the modal frequencies or the subset with the highest Fisher ratio, shows that both populations (accelerograms before and after retrofitting) can be clearly differentiated, despite some overlap in samples from both classes. A simple K-means algorithm for two clusters provides an accuracy of 70%, meaning that 7 out of every 10 samples are classified correctly. The analysis of the confusion matrix for this binary classification problem shows that only 1.39% of the samples classified as before retrofitting are misclassified, whereas for samples classified as after retrofitting, this percentage rises to 56.09%.

The goal of a more sophisticated classifier is to reduce the percentage of misclassified samples, particularly the 56.09% of after retrofitting samples that are classified as before retrofitting, indicating unsuccessful retrofitting in these cases. To achieve this, we used a random forest (RF) algorithm, which, on various 75/25 bags constructed randomly, provides a median accuracy of 99.19%. RF is an ensemble learning algorithm used for classification and regression tasks, and its value lies in analyzing the uncertainty of predictions.

By selecting the best RF model from the 100 simulations conducted and testing it on the entire dataset, an accuracy of 99.77% is achieved, with a nearly perfect receiver operating characteristic (ROC) curve. The analysis of the confusion matrix shows that only 0.12% of the samples classified as before retrofitting are misclassified, while for samples classified as after retrofitting, this percentage drops to 0.32%. Although the decision tree was trained using only a portion of the dataset, this final evaluation on the entire dataset offers insights into the model's predictive ability. However, this practice can overestimate the model's predictive power, so these final results should be interpreted with caution.

Once trained, this classifier can be stored and applied to new samples from the retrofitting of other civil structures, allowing for an estimate of retrofitting success and its probability.

2. Materials and methods

The flowchart of the methodology followed in this paper is shown in Figure 1. It includes three steps: Data Preprocessing, Visual and Discriminatory Analysis, and Final Classification, which are explained in the following subsections.

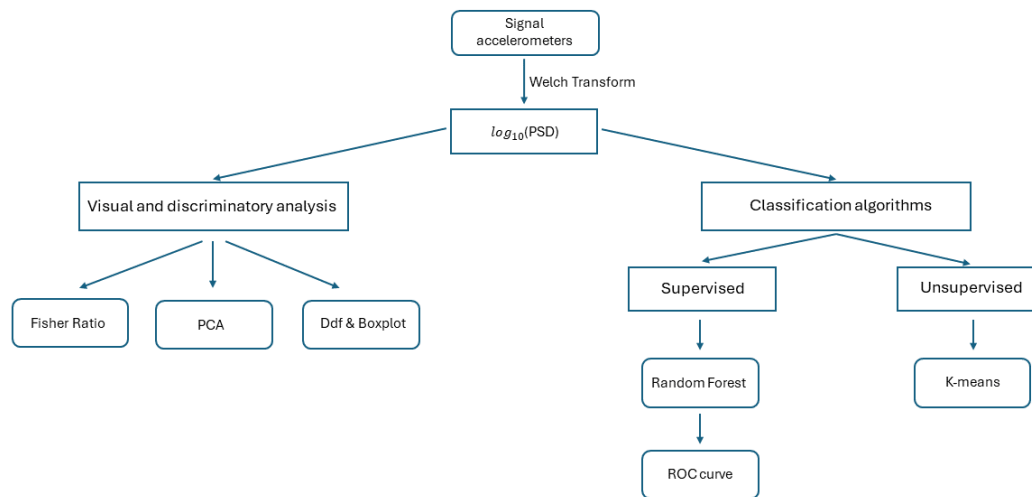


Figure 1. Flow chart of the methodology.

2.1. The data

The KW51 bridge is a steel bowstring bridge located in Leuven, Belgium (latitude 50.9004 N, longitude 4.7066 E), spanning a length of 115 m with a width of 12.4 m. It supports two railway tracks connecting the stations of Herent (2.3 km away) and Leuven (2.2 km away), with a maximum enforced speed of 160 km/h. The bridge serves passenger trains and was opened to traffic in 2003.

An inspection revealed the need for a retrofitting plan to address a construction-related issue. Bolted joints connecting the bridge deck and arches were reinforced by welding steel boxes between the structural components to ensure the bridge's safety and stability. Long-term monitoring occurred from October 2, 2018, to January 15, 2020, with retrofitting taking place from May 15 to September 27, 2019. A monitoring system with 12 accelerometers was installed along the bridge to measure its dynamic behavior, alongside other sensors positioned to collect environmental parameters such as temperature and relative humidity. Datasets were collected over three periods: (i) before retrofitting, from October 2, 2018 to May 15, 2019; (ii) during retrofitting, from May 15, 2019 to September 27, 2019; and (iii) after retrofitting, from September 27, 2019 to January 15, 2020. For each operational day, vibration data (accelerations, strains, and displacements) corresponding to two train passages were recorded within a time window starting 10 seconds before the train entered the bridge and ending 30 seconds after it exited. Additionally, temperature and relative humidity data were included. In cases where a specific sensor was absent or malfunctioning, numerical values were replaced with "NaN" [17].

The environmental data collected, hereinafter referred to as ambient data, will be considered for both the before and after retrofitting stages, based on the aforementioned dates. The available data,

captured by accelerometers placed in various parts of the bridge, measure vibrations and movements caused by environmental conditions. These devices detect changes in acceleration at different points of the bridge and record this information. There are a total of n records, each containing m data points. It is assumed that, before the retrofitting, structural damage was present and that it was remedied by the retrofitting process; thus, the Before-Ambient data reflects a damaged structure, while the After-Ambient data represents a healthy structure.

Figure 2 displays the acceleration data from one of the twelve accelerometers before and after retrofitting. The key observation regarding both signals is that, after retrofitting, the signal amplitude slightly decreases and becomes less homogeneous in value. Some data points exhibit amplitudes higher than the average. In both cases, before and after retrofitting, the number of measurements collected by all accelerometers is 247,742. This large volume of data makes the problem highly indeterminate, as the measurements are also contaminated with noise.

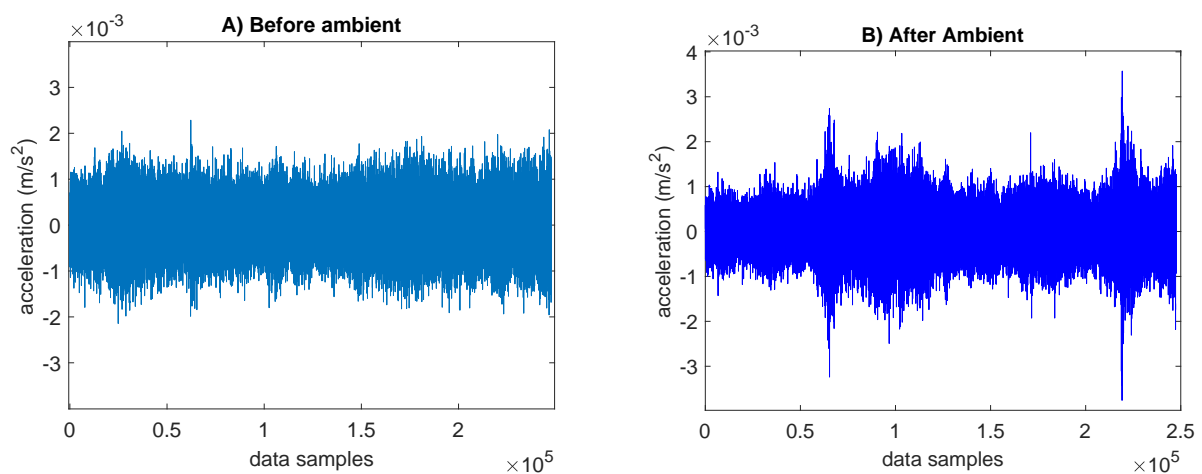


Figure 2. A) Data collected by an accelerometer before retrofitting. B) Data collected by an accelerometer after retrofitting.

2.2. Dimensionality reduction via the Welch method

To address the high dimensionality in the time domain, the data is analyzed in the spectral domain using PSD estimation via the Welch method [18]. This approach allows the analysis of accelerometer signals in the spectral domain, as the PSD provides a detailed breakdown of each frequency's contribution to the signal's overall power, enabling the identification of key frequencies in the accelerometer data. Frequency distributions often reveal distinctive features that can differentiate between various signal classes or extract valuable information. Different signal types have unique frequency profiles, and analyzing these profiles helps in recognizing and characterizing the signals. In engineering systems like bridges, faults or anomalies typically manifest as changes in the signal's frequency content. Therefore, examining these frequency distributions is crucial for fault detection, diagnosis, and proactive maintenance [19, 20].

To estimate the PSD of the accelerometer signals, we employ the Welch method, which can be interpreted as a windowed Fourier transform. This approach involves dividing the signal into overlapping segments, computing a periodogram for each segment, and averaging these periodograms

to obtain the final PSD estimate. Additionally, a window function is applied to each segment in the time domain to minimize edge effects and improve the accuracy of the analysis.

Welch's method is particularly advantageous in the context of SHM due to its ability to handle nonstationary signals and reduce computational requirements. The steps of the Welch method are as follows:

(1) Divide the signal $X(j)$ of length N into K_{max} overlapping segments, each of length L , with a step size h . This can be described as:

$$X_k(j) = X(j + (k - 1)h), \quad j = 0, 1, \dots, L - 1, \quad k = 1, 2, \dots, K_{max},$$

where $K_{max} = \frac{N-L}{h} + 1$. This process is called segmentation.

(2) Apply a window function $W(j)$ to each segment to reduce spectral leakage:

$$X_k^w(j) = X_k(j) \cdot W(j), \quad j = 0, 1, \dots, L - 1.$$

This task is called windowing.

(3) Compute the discrete Fourier transform (DFT) of each windowed segment:

$$A_k(n) = \sum_{j=0}^{L-1} X_k^w(j) e^{-i2\pi \frac{nj}{L}}, \quad n = 0, 1, \dots, L - 1.$$

This task is called spectral transform.

(4) Calculate the periodogram for each segment (f_n frequency) and average them to obtain $\hat{P}(f_n)$:

$$P_k(f_n) = \frac{1}{L} |A_k(n)|^2, \quad f_n = \frac{n}{L}, \quad n = 0, 1, \dots, \frac{L}{2},$$

$$\hat{P}(f_n) = \frac{1}{K_{max}} \sum_{k=1}^{K_{max}} P_k(f_n).$$

This task is called PSD estimation.

This method offers significant advantages in SHM, as it can handle nonstationary signals with low computational requirements.

Throughout this article, we refer to modal frequencies, which are fundamental characteristics of dynamic systems and correspond to the frequencies at which a system naturally vibrates in the absence of continuous external forces. These frequencies are intrinsic to the system's structure and depend on its physical properties.

The order of magnitude of the estimated PSD, $\log_{10}(PSD)(\mathbf{f})$, at the modal frequencies will be used as decision variables in the binary classification problem related to retrofitting.

In the modeling, it is assumed that structural damage existed before retrofitting, which necessitated the relevant repair. The data obtained after retrofitting is considered signals collected in the absence of structural damage.

2.3. Discriminatory analysis of the modal frequencies

To assess the efficacy of retrofitting, the Fisher ratio serves as a robust tool for visualizing significant changes in vibration patterns, allowing us to discern whether the observed differences in accelerometer data are due to the retrofitting process or might arise from random effects.

The Fisher ratio provides a measure of the discriminatory power of an attribute in a binary classification problem. The Fisher ratio of the attribute j can be expressed as follows:

$$FR_j = \frac{(\mu_{1j} - \mu_{2j})^2}{\sigma_{1j}^2 + \sigma_{2j}^2}, \quad (2.1)$$

where μ_{1j} y μ_{2j} are the means of the attribute j within the classes 1 and 2, and σ_{1j}^2 y σ_{2j}^2 are their respective variances.

The Fisher ratio represents the individual prior discriminatory power of each attribute. Attributes with the highest Fisher ratios enhance the classification of low-frequency components, while those with the lowest Fisher ratios help explain the finer details of the discrimination. In our case, the attribute $\log_{10}(PSD)(\mathbf{f})$ sampled in the vector of modal frequencies of the bridge will be used. The most discriminative frequencies are those with a high Fisher ratio, indicating high between-class variance and low within-class variance. Naturally, the individual discriminatory power of variables is altered when they work in synergy, as is the case with most classifiers.

An alternative perspective is obtained through a boxplot. In each box, the middle mark indicates the median, while the lower and upper ends of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that are not considered outliers, and outliers are individually represented by a cross.

Both analyses will be used in the results section to examine the discriminatory power of the different modal frequencies in the retrofitting problem.

2.4. Principal component analysis

For visual analysis of the separability between both classes (before and after retrofitting), we employ PCA to project the data into lower-dimensional PCA subspaces of dimensions two and three [21]. To perform PCA analysis, the data is organized in a matrix

$$X \in M_{ns \times nf},$$

which consists of the $\log_{10}(PSD)(\mathbf{f})$ measurements taken at the 14 modal frequencies, where ns is the number of samples and nf is the number of modal frequencies.

Mathematically, PCA is a linear transformation in the space of predictor variables calculated from the centered covariance matrix C , which is symmetric and positive semi-definite:

$$C = (X - \mu)^T (X - \mu) = Z^T Z \in M_{nf \times nf},$$

where μ represents the mean of the data, and $Z = X - \mu$. The objective is to find a subspace where the projection captures the majority of the variability observed in matrix X .

As explained, the covariance matrix C allows orthogonal diagonalization, yielding eigenvalues and eigenvectors given by the expression:

$$C = VDV^T, \quad D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{14} \end{pmatrix}, \quad V = [\mathbf{v}_1 \dots \mathbf{v}_{14}].$$

Here, D is a diagonal matrix containing the eigenvalues λ_i , and V is the matrix whose columns are the eigenvectors \mathbf{v}_i , with the number of eigenvectors corresponding to the number of modal frequencies. The matrix V thus provides an orthogonal basis for the modal frequency space.

Next, the eigenvalues are sorted in descending order, and the corresponding eigenvectors are arranged accordingly. The k largest eigenvalues are selected, and the corresponding eigenvectors form the principal components. Finally, the data contained in matrix Z is projected onto the new subspace defined by the top k eigenvectors using:

$$T = ZV_k,$$

where T is the transformed data matrix with reduced dimensionality, and V_k is the matrix whose columns are the k eigenvectors corresponding to the k largest eigenvalues.

2.5. *K-means classifier*

In this context, K-means is used as a baseline algorithm to cluster and classify the $\log_{10}(PSD)(\mathbf{f})$ data before and after retrofitting, with the goal of identifying clusters that represent the normal behavior of the bridge and those that indicate potential damage or wear.

This algorithm is initially unsupervised, as the calculation of centroids does not rely on class information. Class labels are only used when determining the final accuracy of the classifier.

K-means is straightforward to implement, and the tuning parameters are as follows:

- (1) Number of clusters (N_c): This represents the number of groups into which the observed data will be divided. In this case, $N_c = 2$.
- (2) Number of initializations (N_i): This represents the number of times the clustering is repeated with new initial centroid positions. In this case, $N_i = 5$.
- (3) Maximum number of iterations: This is the maximum number of iterations the algorithm is allowed for a single run (set to 100 in this case).
- (4) Distance norm: The L_1 norm is used in this case, as it produces more robust results. The algorithm's outputs include a matrix containing the final locations of the centroids and a vector of predicted cluster indices ranging from 1 to N_c .

To calculate the accuracy, we consider two cases:

- Acc_1 : The accuracy obtained when class 1 from the K-means algorithm matches the observed class 1.
- Acc_2 : The accuracy obtained when class 1 from the K-means algorithm matches the observed class 2.
- Ultimately, the class assignment that provides the highest accuracy is adopted, that is:

$$Acc = \max(Acc_1, Acc_2).$$

The final classification is performed as described above.

2.6. RF classifier

The objective of this section is to present, within the universe of classification algorithms, one that improves upon the results obtained using the baseline K-means algorithm. The no free-lunch theorem [22] in ML indicates that no algorithm is superior to others when evaluated across the entire set of possible problems. Considering this principle, we have chosen the RF algorithm due to its versatility and robustness, as it operates as a consensus algorithm. Therefore, the objective of this work is not to compare the performance of different algorithms on this problem but to design a unique and robust methodology for addressing the classification of structural damage or retrofitting.

RF is an ML method that excels in classifying data and is particularly effective due to its ability to handle large datasets and multiple features. It operates by constructing a multitude of decision trees during training, with each tree being built by selecting a subset of the features and determining the best split at each node. The final prediction is generated through majority voting among the individual trees [23–26].

To effectively use the RF algorithm, several parameters must be considered. First, the number of trees to be constructed in the model. A larger number of trees generally improves the model's accuracy but also increases computation time and memory usage. In our case, we used 30 trees since a higher number did not yield better accuracy. The training matrix consists of rows representing observations and columns representing features (predictor variables). In this case, there are 14 predictor variables, which are the modal frequencies, and the data is $\log_{10}(PSD)(\mathbf{f})$. Additionally, there is the class label vector, where each entry indicates the class corresponding to each observation in the training matrix.

Some important parameters of the RF algorithm are:

- (1) NumPredictorsToSample: The number of predictor variables to select at random for each decision split, specified as a positive integer. By default, this property is the square root of the total number of variables (which is four in this case).
- (2) Bootstrap Sampling (bootstrap): By enabling bootstrap sampling, each tree is trained on a random subset of the data with replacement. This method, known as bagging, enhances the model's robustness and accuracy by reducing variance. The default value is 1.
- (3) MinLeafSize: The minimum number of observations at a leaf node. Each leaf must contain at least MinLeafSize observations. This parameter affects the minimum number of samples required to split an internal node in the RF algorithm. By default, MinLeafSize is set to 1.
- (4) Maximum depth of Trees: By default, the trees are grown until each leaf node contains fewer than the minimum number of observations required for a split or until the leaves are pure. This procedure is essential for preventing overfitting.
- (5) Out-of-Bag (OOB) prediction: This enables the identification of the most influential features in making predictions, guiding further feature selection and contributing to model refinement. In this case, the adopted value is 1.

2.7. OOB feature importance algorithm

A technique based on the OOB method has been implemented to calculate feature importance using RF. The OOB method utilizes subsets of data not included in the training sample to validate the model, providing an internal estimate of error and eliminating the need for a separate validation set.

The OOB algorithm is as follows:

- (1) A bagging procedure is defined consisting of $N = 100$ independent simulations. For each simulation, the dataset is randomly split into training (75%) and testing (25%) sets.
- (2) For each simulation, k , an RF model with 30 trees is trained using only the training set.
- (3) Using this model, predictions are made on the test set for the current simulation, and the baseline performance is obtained, denoted as BP_k .
- (4) Feature importance is calculated using the test dataset as follows:
 - (a) For each attribute, j , its values are randomly permuted within the test set, and the permuted performance PP_j^k is calculated.
 - (b) The feature importance of attribute j in simulation k is defined as follows:

$$f_i^k(j) = \frac{|BP_k - PP_j^k|}{BP_k} \times 100.$$

- (c) The final importance of attribute j is defined as

$$f_i(j) = \text{median}\{f_i^k(j)\}_{k=1,\dots,N}.$$

Therefore, feature importance is calculated as the relative decrease in model accuracy after permuting the feature, normalized by the model's baseline accuracy. Feature importance is visualized using a bar chart that displays the median importance of each feature across all simulations.

Although the OOB algorithm is, in principle, applicable to any classifier, the combination of cross-validation and OOB methods provides a robust and reliable assessment of the RF classifier. This procedure can also be used to optimize the construction of the final RF classifier.

3. Results and discussion

For the presentation of the results, the same scheme as in the methodology has been followed.

3.1. The data

The data from accelerometers, referred to as "Ambient", is obtained through monitoring the bridge without any external excitation, i.e., under ambient conditions. A similar dataset exists for train passages, labeled as "Train". In this article, only the ambient data is analyzed, which, theoretically, poses a greater challenge for distinguishing responses before and after retrofitting.

The first step is the preprocessing of the data collected by the accelerometers to calculate the PSD at the 14 modal frequencies described in [8]. The processing of the PSD values is complex due to the amount of information they contain and the need to identify significant patterns within this data. Additionally, since these values are very small, their common logarithm has been calculated to facilitate handling and to help identify patterns and trends that might not be visible on a linear scale. Therefore, the analyzed data is $\log_{10}(PSD)(\mathbf{f})$, where \mathbf{f} is the array of frequencies.

The binary classification problem addresses two classes: class 1, representing ambient data before retrofitting, and class 2, representing ambient data after retrofitting. In the Ambient dataset, there are 14,864 examples in class 1 and 16,272 in class 2. Initially, the power spectrum of the 14 modal frequencies of the bridge is analyzed. Consequently, the values of $\log_{10}(PSD)(\mathbf{f})$ correspond to matrices sized $14,864 \times 14$ in class 1 and $16,272 \times 14$ in class 2.

Figure 3 shows the data matrix, where the transition between the two classes can be easily observed. The objective of this analysis is to design an artificial intelligence (AI) that optimizes this distinction.

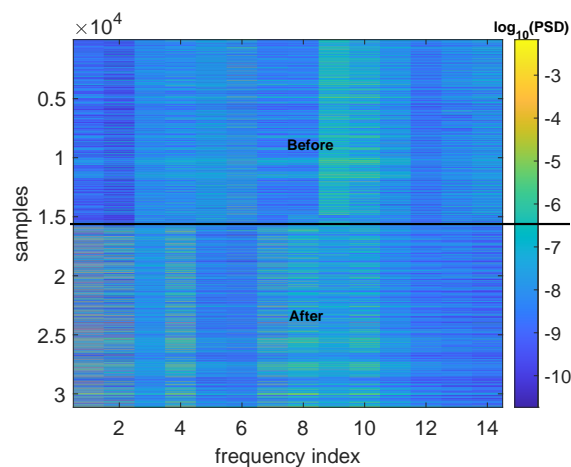


Figure 3. Data matrix ($\log_{10}(PSD)(\mathbf{f})$) for the two classification classes. Each row of the image represents the $\log_{10}(PSD)(\mathbf{f})$ values for the 14 modal frequencies. The upper section contains samples taken before retrofitting, while the lower section contains samples taken after retrofitting. The black line approximately separates the two classes. In this case, spectral analysis is performed exclusively on the 14 modal frequencies.

3.2. Discriminatory analysis of the modal frequencies

Regarding the discriminatory analysis of the attributes (modal frequencies), Figure 4 shows the Fisher ratio values for the Before-Ambient and After-Ambient PSD classes for each of the 14 modal frequencies. It can be observed that only five modal frequencies have a Fisher ratio greater than 0.2, and the most discriminatory modal frequencies correspond to low (0.51, 1.24 Hz) and mid-range frequencies (3.55, 3.88, 4.11 Hz). Low values of the Fisher ratio indicate that the variability between the two groups, $\log_{10}(PSD)(\mathbf{f})$ before and after retrofitting, is small compared to the variability within them. As shown in Figure 4, the modal frequencies with values of 1.89, 4.3, 4.89, 5.33, 6.33, and 6.88 (corresponding to the indexes 3, 10, 11, 12, 13, and 14) do not discriminate between damage and no damage, that is, between before and after retrofitting.

Figure 5 displays the Fisher ratio values in ascending order of frequencies for the Before-Ambient and After-Ambient classes, calculated using 100 frequencies evenly distributed in the interval [0.05, 7], covering the range of modal frequencies. It is notable that the curve exhibits a distribution of Fisher ratio values with a similar profile across groups, as seen in Figure 4. Fisher ratios greater than the mean are represented in yellow, while the rest are shown in red.

Upon comparing both graphs (Figures 4 and 5), it becomes evident that in the case of the 100 frequencies, there are groups with Fisher ratios above the mean (frequency interval [5.52, 5.95]) that are not visible in the graph corresponding to the modal frequencies. Additionally, Figure 5 shows Fisher ratios below the mean interspersed within different frequency intervals, such as [0.05, 0.12] and [1.6, 1.9], which are not represented in the graph corresponding to the modal frequencies.

As expected, the set of 100 frequencies provides a more detailed representation of the Fisher ratio

within each subinterval due to the finer frequency sampling. In conclusion, the spectral analysis with the 100 evenly distributed frequencies does not yield different results from those obtained with the 14 modal frequencies.

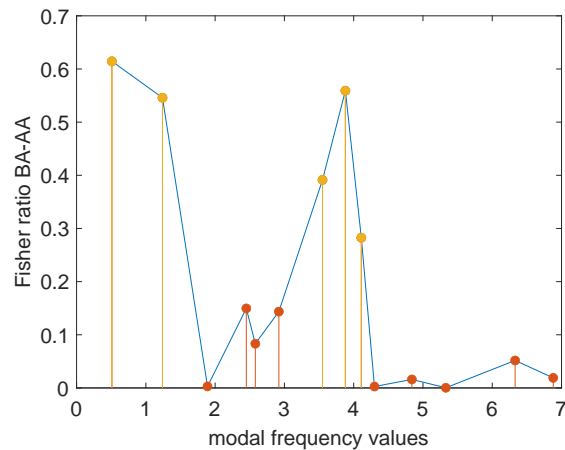


Figure 4. Fisher ratio for $\log_{10}(PSD)(\mathbf{f})$ before and after retrofitting as a function of modal frequency values. Modal frequencies with higher Fisher ratio values are regarded as more discriminatory for this classification problem.

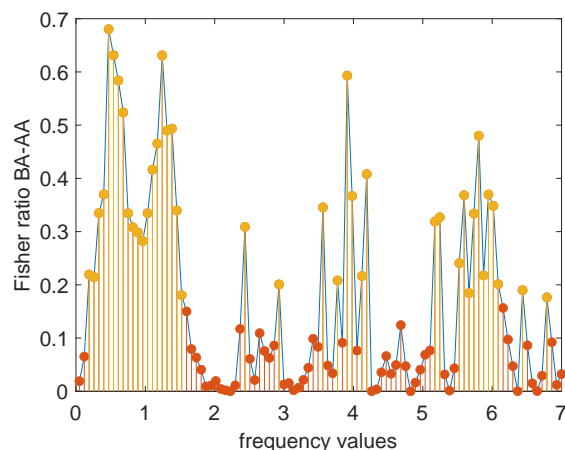


Figure 5. Fisher ratio for $\log_{10}(PSD)(\mathbf{f})$, before and after retrofitting, using 100 equally distributed frequencies referenced in the text. This figure offers greater granularity of the discriminatory power across different frequencies compared to Figure 4.

To better understand the discriminatory power of the 14 modal frequencies using $\log_{10}(PSD)(\mathbf{f})$ as an attribute, Figure 6 displays histograms of this attribute for the different frequencies in both classes. As expected, a greater overlap and/or higher variability of distributions can be observed for the less discriminative frequencies. Figure 7 shows the respective boxplots before and after retrofitting. The boxplots indicate that the median values of $\log_{10}(PSD)(\mathbf{f})$ exhibit a narrower range in the case of “after retrofitting” compared to “before retrofitting”. This observation aligns with the initial accelerometer signals, suggesting that the retrofitting operation was successful in reducing vibrations within the structure. In both cases, anomalous values are more frequent for the larger PSD values.

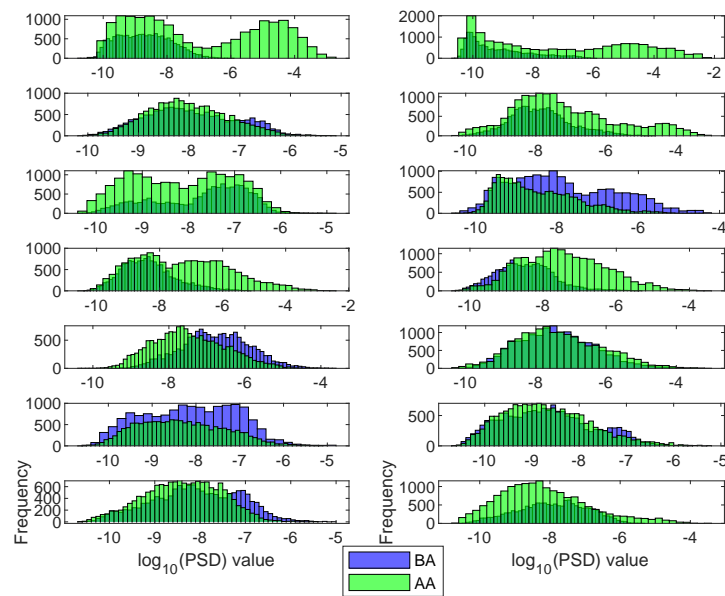


Figure 6. Combined histogram of $\log_{10}(PSD)(f)$ values for both classes (before and after retrofitting) at the 14 modal frequencies. These histograms are arranged in ascending order of the modal frequency values, from left to right and top to bottom. When the histograms overlap or coincide, the corresponding modal frequency exhibits lower discriminatory power for the classification problem related to retrofitting. The content of this figure is directly related to the Fisher ratio shown in Figure 4.

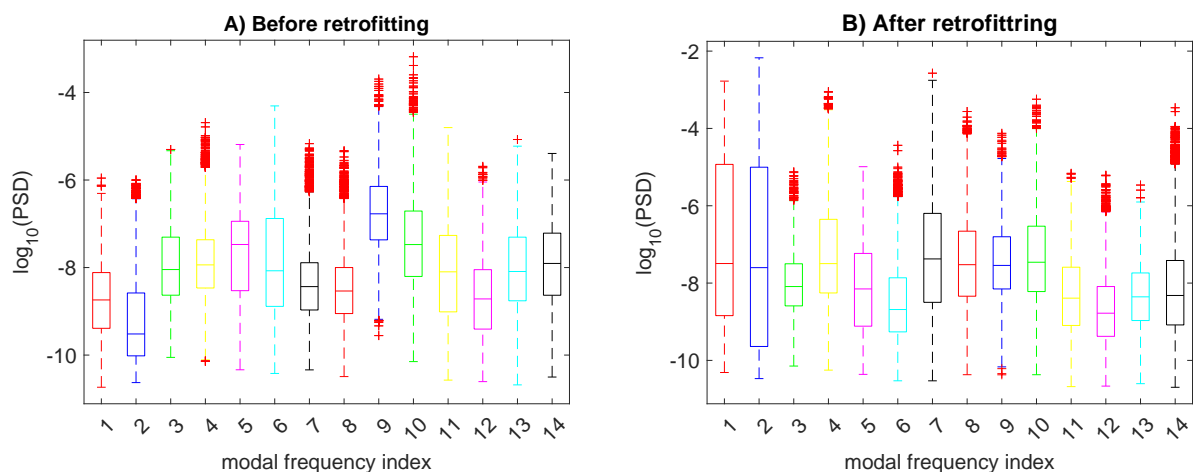


Figure 7. Boxplots of $\log_{10}(PSD)(f)$ obtained from data before (A) and after (B) retrofitting for each modal frequency. The box represents the interquartile range (IQR), encompassing the middle 50% of the data. The bottom and top edges of the box indicate the first quartile (Q1) and the third quartile (Q3), respectively. The line inside the box represents the median (Q2), which is the middle value of the dataset. The whiskers represent anomalous values of the $\log_{10}(PSD)(f)$ distribution in each modal frequency.

3.3. Principal component analysis

Figure 8 shows the projection of the data ($\log_{10}(PSD)(\mathbf{f})$) onto the 2D and 3D PCA spaces sampled at the three most discriminatory modal frequencies: 0.51, 3.88, and 1.24 Hz. It is notable that the first two principal components form two distinct groups with some overlap. Although the data is not linearly separable, there is some clear organization, as the pre-retrofitting data predominantly exhibits negative PC1 values. It is worth mentioning that the first principal component explains 53.65% of the total variance, while the first two principal components account for 77.45%. To surpass 90% of the total variance (91.75%), the first five principal components are needed. Figure 9 shows the cumulative energy curve of this analysis. These figures provide insight into the discriminatory capability of the 14 modal frequencies, as they can be reduced to a lower-dimensional space without compromising their discriminatory power.

Once the most discriminative modal frequencies have been determined, different algorithms are used to solve the binary classification problem in the spectral domain, showing the confusion matrix analysis and the ROC curve of this problem obtained using an RF algorithm.

Following the same reasoning and using the K-means algorithm as a baseline, an accuracy of 70.87% was achieved by training it with the five most discriminatory modal frequencies. The confusion matrix for this classifier was

$$C = \begin{pmatrix} TB & FA \\ FB & TA \end{pmatrix} = \begin{pmatrix} 14760 & 104 \\ 8965 & 7307 \end{pmatrix},$$

where TB represents the number of samples correctly classified as before retrofitting; FA represents the number of samples incorrectly classified as after retrofitting; TA represents the number of samples correctly classified as after retrofitting; and FB represents the number of samples incorrectly classified as before retrofitting.

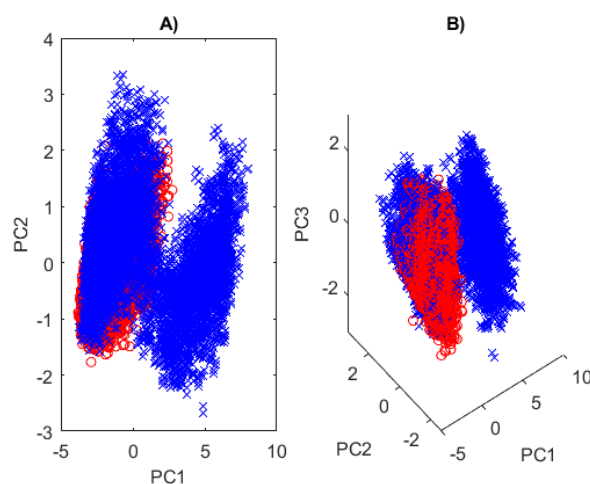


Figure 8. A) Projection of the data ($\log_{10}(PSD)(\mathbf{f})$) onto the 2D PCA space for both classes: before retrofitting (red) and after retrofitting (blue). B) Projection onto the 3D PCA space, with the same color scheme. The data is not linearly separable, although there is some clear organization, as the pre-retrofitting data predominantly exhibits negative PC1 values.

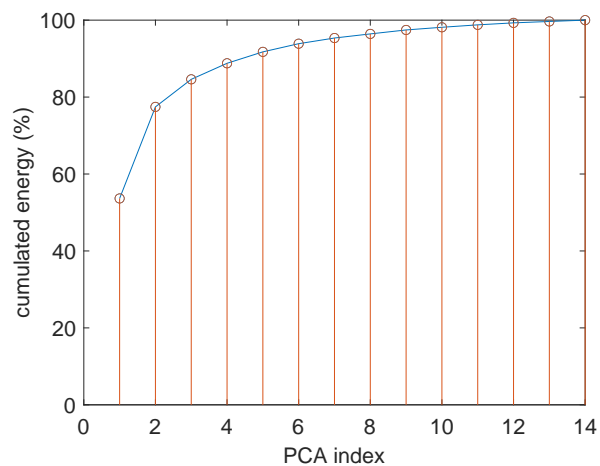


Figure 9. Cumulated energy curve for the PCA analysis. In this curve, it can be seen that the first five principal components capture more than 90% of the variability observed in the matrix X of observations, which contains $\log_{10}(PSD)(\mathbf{f})$ for the modal frequencies of the different accelerograms (before and after retrofitting).

3.4. K-means classifier

To understand the misclassification, Figure 10 shows the projection of the real data and the K-means classification onto the first two spectral mode coordinates. It can be observed that this simple classifier cannot accurately account for the *FB* group, which consists of samples from the “after retrofitting” group that are misclassified as “before retrofitting”. This misclassification occurs because K-means is a linear classifier and is unable to effectively distinguish the overlapping regions.

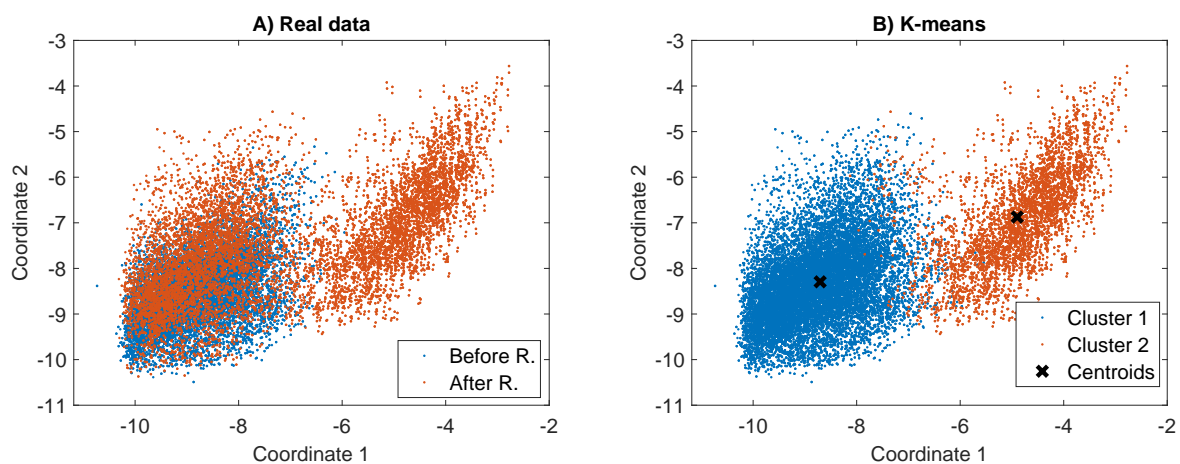


Figure 10. A) Real data projected onto the two first coordinates (spectral frequencies). In this case no previous PCA projection is performed. B) Baseline K-means classification using 14 modal frequencies. The projection onto the first two coordinates is used solely for visualization purposes.

Additionally, to better understand the confusion matrix of the K-means classifier, Figure 11 shows the median signatures of $\log_{10}(PSD)(\mathbf{f})$ for the four different groups identified by the confusion matrix for the K-means classifier. It can be observed that the plots for *FB* and *TB* are very similar, indicating that there is $\log_{10}(PSD)(\mathbf{f})$ data corresponding to the “after retrofitting” group that has been misclassified as “before retrofitting”. This suggests that retrofitting has not had the desired effect in those cases.

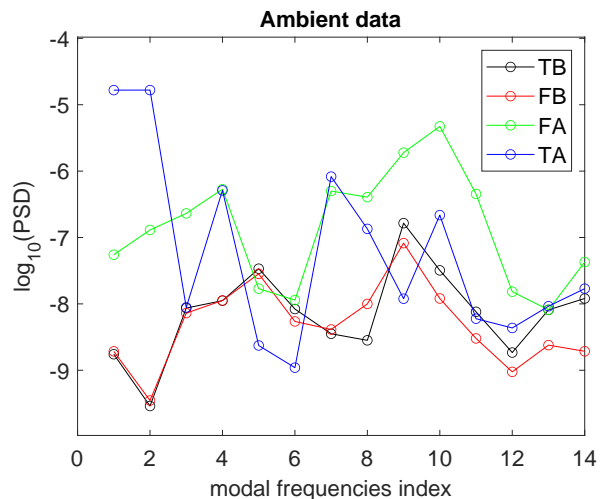


Figure 11. Median $\log_{10}(PSD)(\mathbf{f})$ values obtained from the samples belonging to the different groups of the confusion matrix: A) TB: True before retrofitting (black), B) FB: False before retrofitting (red), C) FA: False after retrofitting (green), and D) TA: True after retrofitting (blue).

3.5. RF classifier

To improve classification accuracy, particularly for the 56.09% of samples incorrectly classified by K-means as “before retrofitting”, we employed an RF algorithm. This algorithm was tested on several 75/25 data splits.

Figure 12 illustrates the cumulative distribution function (CDF) of the accuracy of the RF classifier across 100 independent simulations. This graph is essential for demonstrating the stability and robustness of the classifier’s performance. The x-axis represents accuracy, ranging from 98.8% to 99.6%, while the y-axis shows the probability associated with these accuracy values. The median accuracy, indicated at a probability of 0.5, is approximately 99.2%. This high median value, coupled with the narrow range of accuracy values (low IQR), highlights the classifier’s consistency and reliability when subjected to different random training sets. Such stability is crucial in practical applications, ensuring the classifier performs reliably under various conditions.

The RF algorithm allows for the analysis of the importance of different attributes using a method known as OOB feature importance via a 75/25% cross-validation procedure. OOB samples (25%) are employed to calculate the importance of each feature by measuring the decrease in the model’s accuracy when the values of a particular feature are permuted. Features that cause a significant drop in accuracy are considered more important. Figure 13 shows the OOB feature importance graph. As can be observed, the modal frequencies 3.88, 4.11, 4.30, 4.84, and 6.88 appear to play a more significant

role in the majority voting process. This result is slightly different from the prior discriminatory power of the features given by the Fisher ratio curve (Figure 4). Nevertheless, both graphs agree on the importance of frequencies between 3.88 and 4.84. More surprising is the importance of the frequency 6.88, which, in principle, would explain the higher frequencies in the classification problem (lower Fisher ratio values). This can be explained by the nonlinearity of the RF classifier. In RFs, the different attributes work in synergy, so their discriminatory power differs from what is suggested by their individual Fisher ratios.

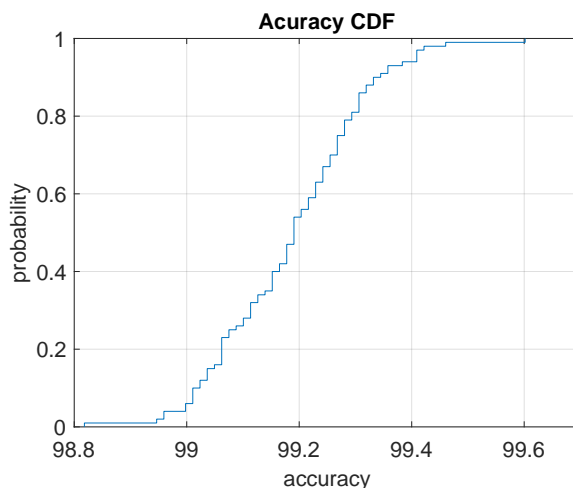


Figure 12. RF classifier. The CDF of the classifier's accuracy was obtained from 100 independent simulations. The median accuracy corresponds to a probability of 0.5, which in this case is approximately 99.2%. The minimum accuracy is 98.8% and the maximum is 99.6%. This graph demonstrates the stability of the classifier's accuracy across different random training sets (cross-validation 75/25).

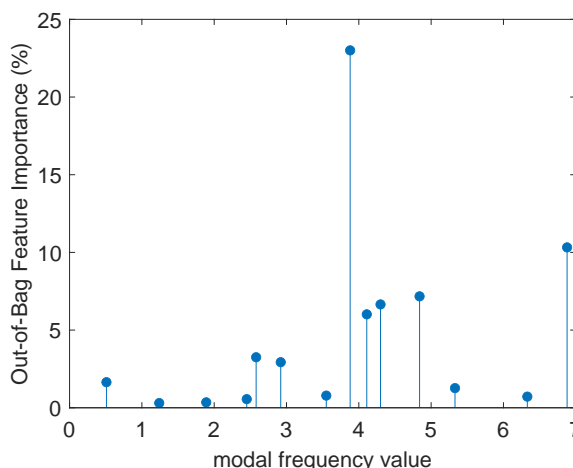


Figure 13. OOB feature importance for the RF classifier. The modal frequencies 3.88, 4.11, 4.30, 4.84, and 6.88 appear to play a more significant role in the majority voting process.

Figure 14 shows the importance plot for 100 equally spaced frequencies. It should be noted that the individual importance of each frequency is lower in the case of the 100 frequencies, as the percentage has been divided by 100 instead of by 14. However, if a cluster of 7 frequencies around 4 Hz is taken—this being the frequency range of greatest importance—a value similar to that of Figure 13 is obtained.

Finally, Figure 15 displays the ROC curve for the RF classifier. This curve illustrates the diagnostic ability of the binary classifier system as its discrimination threshold is varied. Here, the curve resembles an almost perfect unit step function, which indicates near-perfect classification performance. The x-axis represents the false positive rate (FRP) ($1 - \text{Specificity}$), while the y-axis represents the true positive rate (TPR) (Sensitivity). An ROC curve that hugs the top-left corner of the plot signifies high sensitivity and specificity, meaning the classifier has a high true positive rate and a low false positive rate. The almost ideal shape of this curve further validates the classifier's exceptional performance, indicating that it can distinguish between the two classes with minimal errors.

In an ideal model, the TPR would be 1, and the FPR would be 0, meaning it correctly classifies all positive instances and does not incorrectly classify any negative instances as positive. The utility of the ROC curve lies in its ability to evaluate and compare predictive models without being affected by the imbalance in the classes of the target variable. The area under the ROC curve (AUC-ROC) gives an indication of the model's quality. The closer it is to 1, the better the model performs in terms of separating positive and negative classes. In this case, the results obtained are Accuracy: 99.84% and AUC: 0.9999, reflecting the high classification accuracy of the proposed model.

Additionally, the results obtained with RF using the 100 equidistant frequencies are similar to those obtained with the 14 modal frequencies: accuracy: 99.83% and AUC:0.9999. Therefore, it is concluded that there is no loss of information and precision when using only the modal frequencies. As a result, a much simpler model has been preferred, although both can be used for predictive purposes.

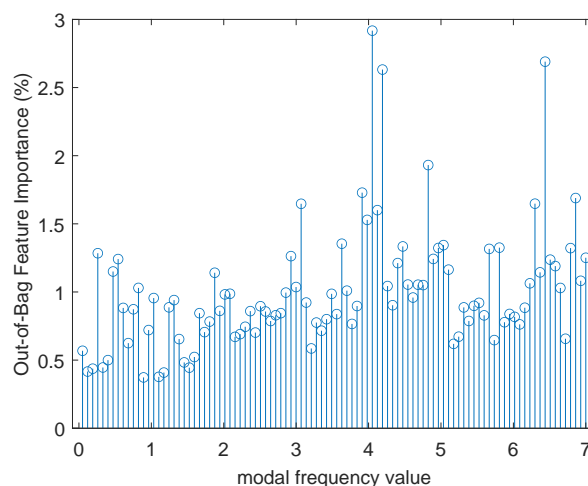


Figure 14. OOB feature importance for the RF classifier using the 100 equidistributed frequencies. The importance shows lower values because it is distributed among a larger number of features.

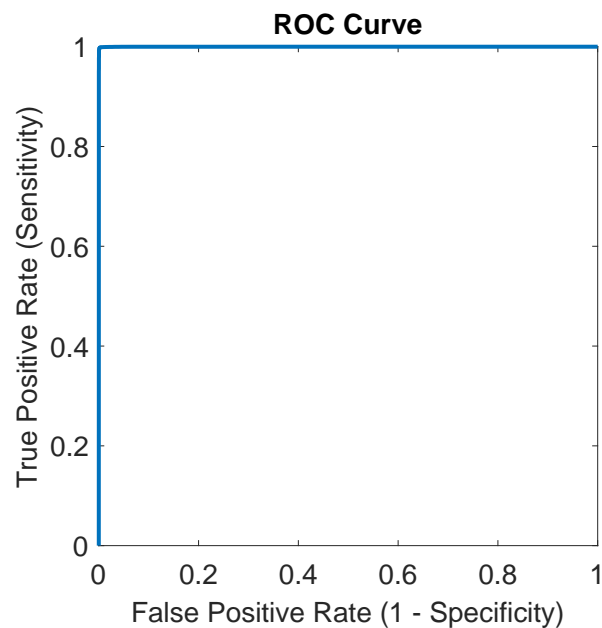


Figure 15. ROC curve for the RF classifier. The curve resembles an almost perfect unit step function, indicating near-perfect classification performance.

4. Conclusions

This research focused on developing a methodology to train an AI system for predicting the impact of retrofitting on civil structures. The paper demonstrates that ML techniques significantly enhance the predictive accuracy of SHM systems. By analyzing large volumes of sensor data, ML models can identify complex, nonlinear patterns indicative of structural damage, often overlooked by human analysts. The application of the Welch transform for dimensionality reduction proved effective in identifying the most predictive features (modal frequencies) for assessing retrofitting effects. This transformation reduced data complexity while preserving critical information necessary for accurate damage prediction.

The baseline K-means algorithm achieved a reasonable classification accuracy of 70% by projecting accelerogram data before and after retrofitting into lower-dimensional spaces. However, a more advanced RF algorithm achieved a significantly higher median accuracy of 99.19%, with excellent ROC curve performance and low misclassification rates. The final model, tested on the entire dataset, achieved an impressive accuracy of 99.77%, underscoring its robustness and reliability in predicting structural health.

Previous discriminatory analysis showed that although the data is not linearly separable, they allow for good differentiation through the PSD of accelerometers estimated at different modal frequencies. Specifically, low and mid-range modal frequencies exhibit higher individual discriminatory power. Subsequently, the RF classifier demonstrated that the most relevant modal frequency in this study was 3.88 Hz. However, all modal frequencies appear to be important since, in the RF algorithm, they work together, slightly differently from what their individual discriminatory power, as indicated by the

Fisher ratio, might suggest.

The high accuracy of the developed classifier indicates its potential utility in real-world applications, such as evaluating the success of retrofitting efforts and informing proactive maintenance strategies. By accurately estimating the condition of new samples, this classifier can help infrastructure managers make informed decisions to prevent structural failures and extend the service life of critical infrastructure. Future research could focus on integrating additional data sources and refining the ML models to further enhance prediction accuracy. Additionally, exploring the deployment of these models in real-time SHM systems could provide continuous monitoring and immediate insights, improving the safety and reliability of civil infrastructure.

Author contributions

A. Presno Vélez, M.Z. Fernández Muñiz and J.L. Fernández Martínez: Conceptualization, Data curation, Visualization, Formal analysis, Investigation, Methodology, Software, Validation, Supervision, Writing – original draft, Writing – review & editing. All authors have read and approved the final version of the manuscript for publication.

Conflict of interest

Prof. M. Z. Fernández Muñiz is the Guest Editor of special issue “New insights of the Application of Inverse Problems and Machine Learning in Science and Technology” for AIMS Mathematics. Prof. M. Z. Fernández Muñiz was not involved in the editorial review and the decision to publish this article. Authors declare no conflicts of interest.

References

1. R. Katam, V. D. K. Pasupuleti, P. Kalapatapu, A review on structural health monitoring: past to present, *Innov. Infrastruct. Solut.*, **8** (2023), 248. <https://doi.org/10.1007/s41062-023-01217-3>
2. O. S. Sonbul, M. Rashid, Algorithms and techniques for the structural health monitoring of bridges: systematic literature review, *Sensors*, **23** (2023), 4230. <https://doi.org/10.3390/s23094230>
3. M. Shibu, K. P. Kumar, V. J. Pillai, H. Murthy, S. Chandra, Structural health monitoring using AI and ML based multimodal sensors data, *Meas. Sens.*, **27** (2023), 100762. <https://doi.org/10.1016/j.measen.2023.100762>
4. Y. J. Cha, R. Ali, J. Lewis, O. Büyüköztürk, Deep learning-based structural health monitoring, *Automat. Constr.*, **161** (2024), 105328. <https://doi.org/10.1016/j.autcon.2024.105328>
5. A. Anjum, M. Hrairi, A. Aabid, N. Yatim, M. Ali, Civil structural health monitoring and machine learning: a comprehensive review, *Fratt. Integr. Strutturale*, **69** (2024), 43–59. <https://doi.org/10.3221/IGF-ESIS.69.04>
6. M. Rodrigues, V. L. Miguéis, C. Felix, C. Rodrigues, Machine learning and cointegration for structural health monitoring of a model under environmental effects, *Expert Syst. Appl.*, **238** (2024), 121739. <https://doi.org/10.1016/j.eswa.2023.121739>

7. H. Son, Y. Jang, S. E. Kim, D. Kim, J. W. Park, Deep learning-based anomaly detection to classify inaccurate data and damaged condition of a cable-stayed bridge, *IEEE Access*, **9** (2021), 124549–124559. <https://doi.org/10.1109/ACCESS.2021.3100419>
8. K. Maes, L. Van Meerbeeck, E. P. B. Reynders, G. Lombaert, Validation of vibration-based structural health monitoring on retrofitted railway bridge KW51, *Mech. Syst. Signal Process.*, **165** (2022), 108380. <https://doi.org/10.1016/j.ymssp.2021.108380>
9. C. R. Farrar, K. Worden, *Structural health monitoring: a machine learning perspective*, John Wiley and Sons, 2012. <https://doi.org/10.1002/9781118443118>
10. C. Scuro, F. Lamonaca, S. Porzio, G. Milani, R. S. Olivito, Internet of Things (IoT) for masonry structural health monitoring (SHM): overview and examples of innovative systems, *Constr. Build. Mater.*, **290** (2021), 123092. <https://doi.org/10.1016/j.conbuildmat.2021.123092>
11. A. Malekloo, E. Ozer, M. AlHamaydeh, M. Girolami, Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights, *Struct. Health Monit.*, **21** (2022), 1906–1955. <https://doi.org/10.1177/14759217211036880>
12. A. Pelle, B. Briseghella, G. Fiorentino, G. F. Giaccu, D. Lavorato, G. Quaranta, et al., Repair of reinforced concrete bridge columns subjected to chloride-induced corrosion with ultra-high performance fiber reinforced concrete, *Struct. Concr.*, **24** (2023), 332–344. <https://doi.org/10.1002/suco.202200555>
13. M. Omori Yano, E. Figueiredo, S. da Silva, A. Cury, I. Moldovan, Transfer learning for structural health monitoring in bridges that underwent retrofitting, *Buildings*, **13** (2023), 2323. <https://doi.org/10.3390/buildings13092323>
14. C. Flexa, W. Gomes, C. Sales, Data normalization in structural health monitoring by means of nonlinear filtering, *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 2019, 204–209. <https://doi.org/10.1109/BRACIS.2019.00044>
15. K. Worden, L. A. Bull, P. Gardner, J. Gosliga, T. J. Rogers, E. J. Cross, et al., A brief introduction to recent developments in population-based structural health monitoring, *Front. Built Environ.*, **6** (2020), 146. <https://doi.org/10.3389/fbuil.2020.00146>
16. P. Gardner, L. A. Bull, N. Dervilis, K. Worden, Domain-adapted Gaussian mixture models for population-based structural health monitoring, *J. Civil Struct. Health Monit.*, **12** (2022), 1343–1353. <https://doi.org/10.1007/s13349-022-00565-5>
17. K. Maes, G. Lombaert, Monitoring railway bridge KW51 before, during, and after retrofitting, *J. Bridge Eng.*, **26** (2021), 04721001. [https://doi.org/10.1061/\(ASCE\)BE.1943-5592.0001668](https://doi.org/10.1061/(ASCE)BE.1943-5592.0001668)
18. P. Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.*, **15** (1967), 70–73. <https://doi.org/10.1109/TAU.1967.1161901>
19. R. Katam, V. D. K. Pasupuleti, P. Kalapatapu, A review on structural health monitoring: past to present, *Innov. Infrastruct. Solut.*, **8** (2023), 248. <https://doi.org/10.1007/s41062-023-01217-3>
20. F. A. Amjad, H. Toozandehjani, Time-Frequency analysis using Stockwell transform with application to guided wave structural health monitoring, *Iran. J. Sci. Technol. Trans. Civ. Eng.*, **47** (2023), 3627–3647. <https://doi.org/10.1007/s40996-023-01224-5>

21. I. T. Joliffe, *Principal component analysis*, 2 Eds., Springer, 2011.
22. D. H. Wolpert, W. G. Macready, No free lunch theorems for optimization, *IEEE Trans. Evolut. Comput.*, **1** (1997), 67–82. <https://doi.org/10.1109/4235.585893>
23. T. K. Ho, Random decision forests, *Proceedings of 3rd International Conference on Document Analysis and Recognition*, **1** (1995), 278–282. <https://doi.org/10.1109/ICDAR.1995.598994>
24. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
25. M. C. Cheng, M. Bonopera, L. J. Leu, Applying random forest algorithm for highway bridge-type prediction in areas with a high seismic risk, *J. Chin. Inst. Eng.*, **47** (2024), 597–610. <https://doi.org/10.1080/02533839.2024.2368464>
26. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2 Eds., Springer-Verlag, 2009. <https://doi.org/10.1007/978-0-387-84858-7>



AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)