Smart Structures and Systems, *Vol. 29, No. 1 (2022) 251-266*
https://doi.org/10.12989/sss.2022.29.1.251

A semi-supervised interpretable machine learning framework for sensor fault detection

Panagiotis Martakis*[1], Artur Movsessian [2a], Yves Reuland [1b], Sai G.S. Pai [3c], Said Quqa [4d], David Garcıa Cava [2e], Dmitri Tcherniak [5f] and Eleni Chatzi [1g]

[1] *Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland*
[2] *School of Engineering, Institute for Infrastructure and Environment, University of Edinburgh, Alexander Graham Bell Building, Thomas Bayes Road, Edinburgh EH9 3FG, UK*
[3] *Intellithink Industrial IoT Labs, Bengaluru, India (previously with Future Cities Laboratory, Singapore ETH Centre, Singapore)*
[4] *Department of Civil, Chemical, Environmental, and Materials Engineering, University of Bologna, Viale del Risorgimento 2, 40136 Bologna, Italy*
[5] *Brüel & Kjaer Sound and Vibration Measurements, Skodsborgvej 307, Naerum 2850, Denmark*

(*Received May 16, 2021, Revised July 30, 2021, Accepted September 8, 2021*)

Abstract. Structural Health Monitoring (SHM) of critical infrastructure comprises a major pillar of maintenance management, shielding public safety and economic sustainability. Although SHM is usually associated with data-driven metrics and thresholds, expert judgement is essential, especially in cases where erroneous predictions can bear casualties or substantial economic loss. Considering that visual inspections are time consuming and potentially subjective, artificial-intelligence tools may be leveraged in order to minimize the inspection effort and provide objective outcomes. In this context, timely detection of sensor malfunctioning is crucial in preventing inaccurate assessment and false alarms. The present work introduces a sensor-fault detection and interpretation framework, based on the well-established support-vector machine scheme for anomaly detection, combined with a coalitional game-theory approach. The proposed framework is implemented in two datasets, provided along the 1st International Project Competition for Structural Health Monitoring (IPC-SHM 2020), comprising acceleration and cable-load measurements from two real cable-stayed bridges. The results demonstrate good predictive performance and highlight the potential for seamless adaption of the algorithm to intrinsically different data domains. For the first time, the term "decision trajectories", originating from the field of cognitive sciences, is introduced and applied in the context of SHM. This provides an intuitive and comprehensive illustration of the impact of individual features, along with an elaboration on feature dependencies that drive individual model predictions. Overall, the proposed framework provides an easy-to-train, application-agnostic and interpretable anomaly detector, which can be integrated into the preprocessing part of various SHM and condition-monitoring applications, offering a first screening of the sensor health prior to further analysis.

Keywords: decision trajectories; decision trajectory assurance criterion; DTAC; interpretable AI; one class classifiers; sensor fault detection; SHAP; SHM

1. Introduction

The built environment and its infrastructure systems form the backbone of modern societies. However, ageing infrastructure components, increasing transportation needs, limited maintenance budgets and the massive carbon footprint of the construction industry are challenging engineers, operators and decision-makers. Structural health monitoring (SHM) has emerged as a powerful enabler of risk-informed extension of the service life of ageing infrastructure and industrial assets by allowing safe utilization of reserve capacity, while reducing excessive safety margins (Smith 2016). Many SHM applications have emerged in the past (An *et al*. 2019, Sohn *et al*. 2001), ranging from data-driven damage detection (Gui *et al*. 2017, Neves *et al*. 2017, Worden *et al*. 2000) to model updating for higher-end damage identification tasks (Jaishi and Ren 2006, Moaveni *et al*. 2009). These include damage characterization and quantification, as well as prognostic tasks, for instance the inference of structural capacity and residual life prediction (Martakis *et al*. 2021, Pai *et al*. 2019, Reuland *et al*. 2017). A common aspect of all SHM applications resides in the reliance on structural sensing to gain insights into structural behavior. Amongst available techniques, vibration-based SHM (Fan and Qiao 2011, Limongelli *et al*. 2016), which relies on the monitoring of dynamic response, is currently the most broadly established and widely researched approach.

In recent years, machine-learning (ML) techniques have

*Corresponding author, Ph.D. Candidate,
E-mail: martakis@ibk.baug.ethz.ch
[a] Ph.D. Candidate, E-mail: artur.movsessian@ed.ac.uk
[b] Ph.D., E-mail: reuland@ibk.baug.ethz.ch
[c] Ph.D., E-mail: saiganesh89@gmail.com
[d] Ph.D. Candidate, E-mail: said.quqa2@unibo.it
[e] Ph.D., E-mail: david.garcia@ed.ac.uk
[f] Ph.D., E-mail: dtcherniak@bksv.com
[g] Ph.D., Professor, E-mail: chatzi@ibk.baug.ethz.ch

ISSN: 1738-1584 (Print), 1738-1991 (Online)

been increasingly applied to vibration data to perform damage detection in engineering structures (Abdeljaber *et al*. 2017, Azimi *et al*. 2020, Bao *et al*. 2019a, Figueiredo and Santos 2018). Such approaches may be classified as supervised, semi-supervised or unsupervised, depending on the data labels required for training. While supervised methods (Tibaduiza *et al*. 2018) capitalize on the availability of labelled data for achieving damage classification, unsupervised approaches have also been proposed and successfully applied in pattern recognition and classification tasks (Tibaduiza *et al*. 2013). For many practical applications, where the availability of labeled data is scarce or tedious to provide, semi-supervised approaches carry promising potential in novelty detection tasks (Bull *et al*. 2018). One-class classifiers (OCCs) trained on data belonging to a known class, namely the "normal" or "healthy" class, gained significant popularity due to fast training and remarkable predictive performance in anomaly detection. Properly trained OCCs attribute new observations either to the initial distribution ("normal") or to an alternative class: outliers / anomalies, a separation that might require a hard definition of a threshold.

In the context of supervised learning, Chen *et al*. (2018) introduced the XGBoost algorithm, achieving particularly rapid learning through parallel and distributed computing, while ensuring efficient memory use. Zhang *et al*. (2018) demonstrated the superior performance of XGBoost classifiers when dealing with multi-dimensional feature sets, especially in terms of preventing model overfitting (Dietterich 1995). Given these benefits, XGBoost has gained popularity across diverse data science applications. Although ML has been successfully applied in damage detection tasks (Figueiredo *et al*. 2011, Long and Buyukozturk 2014, Ying *et al*. 2013), most advanced algorithms fall into the category of black-box models that provide very limited, if any, information regarding the underlying decision-making process. In the aftermath of recent catastrophic failures within the structural and geotechnical engineering context, numerous forensic engineering reports exposed the link between failures and asset-management decisions related to structural components (Alonso *et al*. 2010). Despite recent advances in simulation and computational capabilities, the lack of interpretability of models, be it black-box ML-based or complex finite element models, undermines the trust of practitioners in their predictions and ultimately prevents a broader adoption in real-world problems within the SHM context. This limitation is even more prominent in damage detection and characterization tasks, due to the absence of sufficient labeled data. Moreover, sensor malfunctioning and ageing may lead to erroneous condition assessment (Bao *et al*. 2019b) and redundant sensor placement, which is neither financially nor environmentally sustainable. Particularly to what concerns this latter point, a framework to perform comprehensible and automated detection of sensor malfunctioning that is not structure-specific could not be found in the relevant literature.

The need for techniques to interpret ML model outcomes has delivered rich research in recent literature (Vilone and Longo 2020). Ribeiro *et al*. (2016) proposed the development of interpretable surrogate models, capable of approximating local predictions of an overarching black-box model. Although intuitively simple, this method suffers from the unresolved issue of the neighborhood definition, which bounds the region of fit between surrogate and black-box model. Another approach based on coalitional game theory was recently proposed by Lundberg *et al*. (Lundberg and Lee 2017). The SHAP algorithm allows for the interpretation of individual decisions of black-box models through an efficient computation of the Shapley values, which quantify the contribution of each feature to the overall prediction outcome (Roth 1988, Shapley 1953, Štrumbelj and Kononenko 2014). Shapley values, in the context of ML-based applications, provide insights into the marginal contribution of each feature to the final class prediction (e.g., "normal" or "abnormal"). Therefore, SHAP presents an additive feature attribution method that defines the class output as summation of the real values attributed to each (input) feature. While conventional feature importance algorithms, such as Neighborhood component Analysis (Goldberger *et al*. 2005), estimate the impact in a more global sense, local interpretability allows for a comparison between individual predictions of the black-box model. Assembling this information for each target class uncovers the driving features behind each class prediction, as well as their positive or negative impact. SHAP has been implemented by Lundberg *et al*. within a medical context (Lundberg *et al*. 2018), by Bussmann *et al*. for financial applications (Bussmann *et al*. 2020) and by Parsa *et al*. to derive common causes of accidents (Parsa *et al*. 2020).

Applications of explainable ML in the context of SHM are scarce. Lim and Chi (2019) implemented an XGBoost classification model to estimate the severity of damage levels of bridges, based on visual inspections. Explainable ML enabled the identification of the key features that link to the observed damage. Onchis *et al*. (Onchis and Gillich 2021) combined Local Interpretable Model-agnostic Explanations (LIME) with the SHAP algorithm, in order to infer the location and the depth of cracks in monitored cantilever beams, while emphasizing the contribution of individual features to preventive maintenance decisions. Movsessian *et al*. (2020) introduced a decision tree-based methodology to explain a Mahalanobis distance-based novelty index for damage detection. The methodology was demonstrated on a wind turbine blade with an additional point mass to simulate damage in different locations. Observing the decision process of the decision trees allowed interpreting the occurrence of novelties and ultimately the damage localization. This methodology was further refined by utilizing the SHAP approach and XGBoost decision trees to identify the effect of environmental and operational conditions (EOCs), allowing to differentiate false positives due to temperature effects from true positives triggered by actual damage (Movsessian *et al*. 2021).

In contrast to mechanical components, civil structures are typically unique systems. This undermines any attempt to exploit the scarce-labeled datasets to train generic models that could be applied to other, even similar, structures. To our knowledge, no generic framework for interpretable ML in data-driven sensor fault detection has been proposed to

date.

This contribution aims to merge explainable ML with OCCs that can be trained solely on data acquired during "normal" operation of permanent SHM installations. Exploiting the SHAP algorithm on the evaluations of OCCs enables the tracking of the individual feature importance and the feature dependencies for each abnormality prediction (fault). By implementing the proposed framework in two intrinsically different datasets, it is shown that common sources of abnormal behavior feature similar interpretation patterns. In order to provide an intuitive and compact visualization of the model explanations, the "decision trajectories", originating from the field of cognitive sciences for the visualization of decision landscapes in mouse-tracking experiments (Zgonnikov *et al*. 2017), are defined and applied for the first time in the context of SHM. In this context, decision trajectories demonstrate the accumulation of SHAP values along the feature space to the final SHAP score, which is associated with the model outcome. Decision trajectories may become a powerful tool in fault detection processes by supporting expert-based attribution to a specific class of abnormal behavior, for instance differentiating between sensor and structural failure. Similar visualizations have been demonstrated in the medical field (Athanasiou *et al*. 2020), although no systematic study on the trajectory-patterns has been conducted.

This paper initiates with a description of the proposed interpretable semi-supervised OCC approach and the definition of the decision trajectories (Section 2). The features extracted from monitoring data are presented next (Section 3), with a subsequent description of the case studies, on which the methodology is showcased. Finally, results are presented and discussed, in Section 4.

2. Methodology

The present work combines a semi-supervised framework for sensor fault detection with a recently proposed approach to explain ML-driven decisions, based on the SHAP algorithm. The SVM algorithm is used to identify anomalies, essentially acting as an OCC. The identified anomalies are further used as labels for a supervised classification scheme, namely the gradient boosted decision trees, as materialized in the XGBoost algorithm, which is used to build a relationship between the original features and the three classes, i.e., normal, uncertain and abnormal. Given that the proposed framework relies exclusively on data belonging to a single class (addressed as healthy, reference or normal), it falls into the semi-supervised category, as opposed to supervised approaches that require labels for all types of faults. In a next step, the SHAP values are computed for further

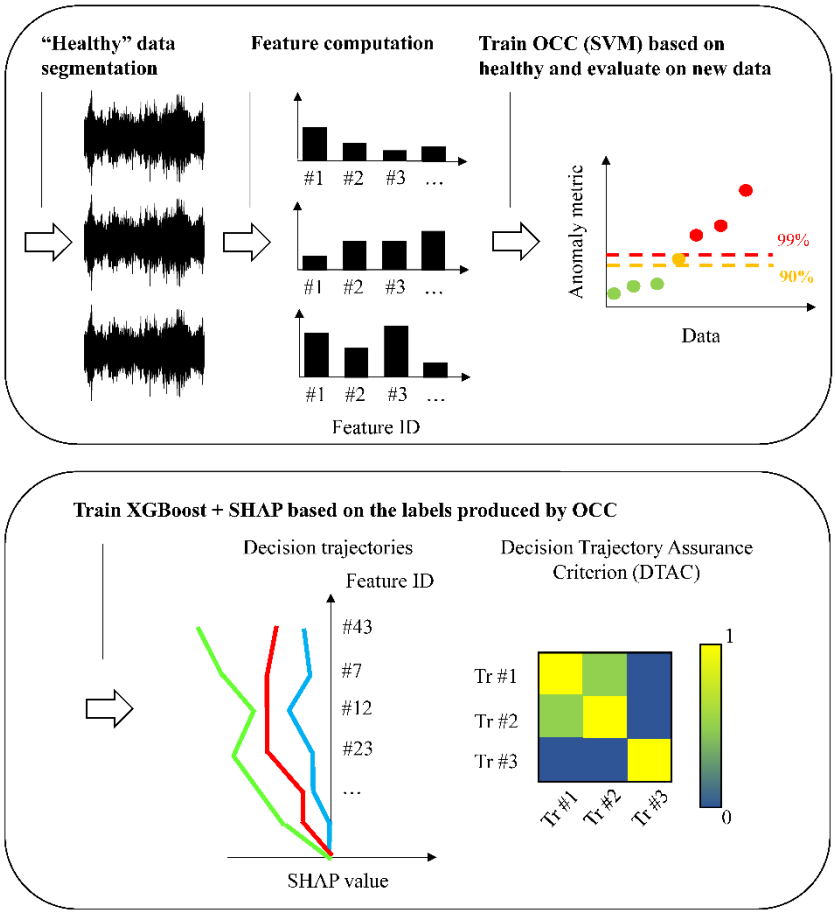


Fig. 1 Illustration of the proposed framework for sensor fault detection (up) and the subsequent interpretation of model decisions (down)

interpreting those results that led to a certain prediction (normal, uncertain or abnormal). The trajectories that are plotted render the computed SHAP values in a hierarchical order of contribution. The proposed framework is illustrated in Fig. 1 and aims to complement conventional one-class classification by supplementing it with a thorough and comprehensive interpretability of model decisions. The framework is designed to be independent of the distinctive structure-specific characteristics of the monitored structure and the type, amount and locations of sensors used, providing an easy to train, project-agnostic tool, which can be seamlessly integrated into the preprocessing part of various SHM applications.

2.1 Preprocessing and feature engineering

The training set comprises "healthy" data in the form of time series, covering a large variety of conditions in which the structural response can be characterised as "normal". Each measuring channel is evaluated independently, allowing for the concatenation of all available data for training. The definition of the reference response is ambiguous, among other reasons due to influences of varying Environmental and Operational Conditions (EOCs) (Avendaño-Valencia *et al*. 2017). In order to account for these effects, which are not linked to damage but to regular operating conditions, the proposed framework can be retrained, as soon as further "normal" data, falling outside the initial training set, becomes available. Nevertheless, for the purposes of this work the reference data, which could be provided by expert judgement, are considered available. The training data are segmented into windows of a predefined length that is meaningful for the studied system. For civil structures with fundamental frequencies above 1 Hz, a minimum length of 60 seconds is suggested. It is not recommended to include any standard preprocessing (i.e., digital filtering, linear trend exclusion and downsampling) at this stage, as such processing may mask potential sensor failures or other anomalies. Subsequently, the segmented data-series are mapped into an extensive feature space, including statistical metrics calculated in time and in frequency domain. A detailed discussion on the implemented features is provided in section 3. All features are normalized, so that the 5th and 95th percentiles of the training set correspond to the limits of the range [−1, 1].

2.2 Training of OCC

The underlying classification task is conducted by means of Support Vector Machines (SVMs) for novelty detection, originally introduced in Schölkopf *et al*. (2000), as an extension of the Support Vector algorithm (Hearst *et al*. 1998) for the case where sufficient "healthy" data are available. SVM classifiers require an initial choice of a kernel function and a scalar parameter to define a delimitation frontier. For the purposes of this project, the radial-basis kernel function (Musavi *et al*. 1992) is applied and a grid search is conducted to define the optimal parameter v, which characterizes the fractions of Support Vectors and outliers and is bounded between zero and one. The SVM-based OCC model is trained to identify the smallest hypersphere comprising all data points of the training set, which allows for the definition of thresholds, above which the value is considered abnormal. In order to account for borderline cases that may be ambiguous, two separate thresholds are defined and correspond to the 90th and the 99th percentile of the training set. Data points that fall between these thresholds are considered uncertain, while predictions above the 99th percentile threshold indicate abnormal behaviour. It is mentioned that these thresholds although case specific, they can be seamlessly defined and adjusted based on expert judgment.

2.3 Training of XGBoost and SHAP

In order to benefit from a computationally efficient implementation of the SHAP algorithm with decision-tree-based models, we train a surrogate XGBoost ensemble to fit the predictions of the OCC. Given a sufficient amount of evaluated data where abnormal behaviour occurs, the predictions of OCC in form of "normal", "uncertain" and "abnormal" consist the labels for the training of the XGBoost classifier. Due to the general scope of the trained classifier, abnormal behaviour may be linked to sensor malfunction, significant changes in EOCs and structural damage or failure. Subsequently, the SHAP algorithm is applied for the interpretation of the XGBoost model decisions. SHAP is an additive feature attribution method, which allows the expression of model decisions as a sum of real values attributed to each feature (Lundberg and Lee 2017). An explanation model α, which is a linear combination of binary features, is defined as follows

$$a(z) = \varphi_0 + \sum_{l=1}^{L} \phi_l z_l \tag{1}$$

where $z_l \in \{0,1\}^L$, L is the length of the feature vector, l the index for a particular feature and $\phi_l \in \mathbb{R}$ is the feature attribution value. The variables z_l represent a feature being observed ($z_l = 1$) or unknown ($z_l = 0$).

The SHAP method defines $f_{\boldsymbol{\chi}}(S) = E[f(\boldsymbol{\chi})|\boldsymbol{\chi}S]$, where $\boldsymbol{\chi}$ represents the feature set and S contains a set of non-zero indexes in z. The feature attribution value for each feature i is computed through the classic Shapley value formulation

$$\phi_i = \sum_{S \subseteq N \setminus \{l\}} \frac{|S|!\,(L - |S| - 1)!}{L!} \left[f_{\boldsymbol{\chi}}\left(S \cup \{l\} - f_{\boldsymbol{\chi}}(S)\right)\right] \tag{2}$$

where N is the set of all input features and S a subset of $N \setminus \{l\}$ withholding feature l.

The SHAP score is calculated by adding the feature attribution values to the baseline φ_0, which comprises the average of all accumulated feature attribution values of the training set. The global importance of feature i in a set of M samples is calculated through the average of the absolute values of the feature attributions φ_i

$$G_i = \frac{1}{M} \sum_{i=1}^{M} |\varphi_i| \tag{3}$$

On a local scale, $\varphi_{i,k}$ expresses the impact of feature i on the model decision k. Feature attributions have either a positive contribution, i.e., pushing the model decision towards the model prediction k, or a negative contribution, which pushes the model decision away from a final prediction k. All feature dependencies that lead to the final decision of the model are encoded in the herein defined decision trajectories (Fig. 1(b)). Starting from the bottom, each prediction line demonstrates how the feature attributions accumulate to the final SHAP score, which is associated with the model decision. In this context, each prediction line illustrates the trajectory of the model decision, implicitly comprising distinctive feature dependencies. Hence, detected anomalies with identical characteristics are expected to yield similar decision trajectories, enabling a better understanding of the origin of faults (root-cause analysis), beyond the limits of the binary one-class classification. The decision trajectories are formed through interpolation of the Shapley attribution values assigned to each feature. In order to evaluate the consistency between different decision trajectories, we propose a measure of Correlation, which is based on the formulation of the well-established Modal Assurance Criterion (MAC), typically used as a quality indicator for modal vectors that are estimated from measured frequency response functions (Allemang 1982). The proposed linearity metric, namely the Decision Trajectory Assurance Criterion (DTAC), compares the first derivative of the evaluated and the reference trajectories and yields a scalar, which expresses the degree of linearity between the two vectors

$$DTAC = \frac{\left|(\boldsymbol{\varphi}'_{eval})^T \boldsymbol{\varphi}'_{ref}\right|^2}{|(\boldsymbol{\varphi}'_{eval})^T \boldsymbol{\varphi}'_{eval}| \left|(\boldsymbol{\varphi}'_{ref})^T \boldsymbol{\varphi}'_{ref}\right|} \quad (4)$$

where $\boldsymbol{\varphi}'_{eval}$ and $\boldsymbol{\varphi}'_{ref}$ refer to the first derivative of the evaluation and reference trajectories respectively.

The DTAC is bounded between zero, indicating no consistent correspondence, and one, indicating a consistent correspondence. Fig. 1(b) comprises a 2D visualisation of the DTAC metric evaluated on three decision trajectories. Assuming that sufficient faults have been classified based on expert judgement, reference trajectories can be defined for each fault type, by utilising a representative statistical metric of all trajectories associated with each specific fault type. To this end, the median is selected, as it is less prone to outliers compared to the statistical mean. The curvature of the decision trajectories contains significant information regarding the feature dependencies and thus the first derivative of the trajectories is implemented in the proposed correlation metric. It is mentioned that the feature order is dictated by the global feature importance from the training set (healthy data) and remains constant.

Illustrating the feature impact on the decisions of black-box models through decision trajectories resembles to the superposition principle, which is applied to complex linear systems, in order to decompose their response into fundamental linear components. The Fourier transform is a characteristic example of such a decomposition. Analysing time-series into monochromatic signals exposes the characteristic frequencies that shape the studied response. Similarly, the decision trajectories decompose black-box model predictions into the impacts of individual features, enabling the detection of key feature-dependencies that characterise individual model decisions. Evaluating decision trajectories of false alarms or "uncertain" predictions can support experts by exposing the misleading features. In addition, similarities in decision trajectories may facilitate associating abnormal signals with specific sources of malfunctioning.

2.4 Summary of the workflow

Overall, the framework for explainable anomaly detection consists of six steps:

i. Compute features from sensor readings collected during different conditions.
ii. Train a SVM as an OCC with a training set composed of exclusively healthy data.
iii. Use the trained OCC to classify new data samples and identify anomalies.
iv. Use the outcome of the OCC as labelled dataset for a supervised classification scheme, namely the gradient boosted decision trees, as materialized in the XGBoost algorithm, which is used to build a relationship between the original features and the three classes, i.e., normal, uncertain and abnormal.
v. Compute the SHAP values to interpret the results from the XGBoost classification, which leads to a deeper understanding of the identified anomalies.
vi. Use the Decision Trajectory Assurance Criterion (DTAC) to identify similarities between trajectories.

3. Features extracted from dynamic data

The selection of an appropriate suite of features that encodes the maximum information from the available data consists the cornerstone of most ML-based applications. In an attempt to keep the proposed framework as generic as possible, without compromising its predictive performance, a comprehensive set of features is designed, including various metrics defined both in time and in frequency domain. All recorded channels are evaluated independently and no correlation metrics are considered (Fig. 1(a)). The selected features are summarized in Table 1 and described in this section. It is noted that the presented feature space can be seamlessly tailored to the studied application.

Descriptive time-domain statistics are used to characterize the statistical distribution of the data samples that form a time-series. Mean, Median and Mode (#1−3) give insights into the central part of the data distribution. Standard-deviation, Variance, Root-Mean-Square and Coefficient of Variation (#4, #5, #8 and #25) are influenced by the amplitude of the "noise" that is present in the signal and by the amplitude of structural oscillations. Finally, Skewness and Kurtosis (#6 − 7) characterize the tails of the sample distribution providing, among others, information about outliers.

Table 1 Overview of the selected time-series features

Feature #	Feature description	Domain
1, 2, 3	Mean, Median, Mode	Time
4, 5, 8, 25	St.-dev., Variance, Coeff. of Variation, RMS	Time
6, 7	Skewness, Kurtosis	Time
9, 10	Relative occurrence frequency of 0 and NaN values	Time
11, 12	Longest time span above and below mean	Time
13	Number of absolute peaks in a fixed time-window*	Time
14	Sum of reoccurring datapoints	Time
16	Percentage of reoccurring datapoints to all datapoints	Time
17	Percentage of unique datapoints to all datapoints	Time
18	Signal auto-correlation considering a fixed lag*	Time
19	Partial auto-correlation considering a fixed lag*	Time
20	Benford correlation value (Hill 1995b, 1995a)	Time
21	Nonlinearity metric c3 (Schreiber and Schmitz 1997)	Time
23	Symmetry boolean value	Time
24	Time reversal asymmetry statistic, defined in (Fulcher and Jones 2014)	Time
28	Sum of absolute values of first derivative	Time
29, 30	Mean value of first and second (numerical) derivative	Time
31	Boolean check if variance is larger than standard deviation	Time
32	Boolean check if standard deviation is larger than 25% of the range of the response distribution	Time
33	Percentage of datapoints that lie beyond 1 standard deviation from the mean	Time
34, 35, 36	St.-dev. of moving averages with short, intermediate and long windows	Time
37, 38, 39	Homogeneity of moving averages with short, intermediate and long windows	Time
42,43	Kurtosis of the FFT and PSD spectra	Frequency
44, 45	Homogeneity of the FFT and PSD spectra	Frequency
46, 48	Central frequency range containing 95% of the FFT and PSD spectra	Frequency
47,49	Central frequency range containing 50% of the FFT and PSD spectra	Frequency

*The length of the fixed time-window/lag is defined as the minimum between data-series length/100 and 3000

Heuristic characterization of the data time-series is used to construct metrics that provide information on the data vector and the quality of the time-series signal. Relative occurrence frequencies of zeroes (#9) and NaN values (#10) may, for instance, be linked to the functionality of the data acquisition system. The longest time-span (number of samples) above and below the mean value (#11 and #12) are other examples of features that are based on heuristics and may be related to the calibration of sensors, drift and connection between sensors and the structure. The number of peaks (#13) in the data series depends on the frequency, but in case of sensor malfunctioning (e.g., square signal) may be significantly reduced.

Derived indicators for dynamic data are indicators that are constructed to reflect dynamic data series. The auto-correlation (#18) or partial auto-correlation with fixed time lags (#19) as defined by Wilson (2016), provide information regarding the periodic repetition of the signal during the selected time window. The Benford correlation value, also referred to as significant-digit law, defined by Hill (1995a, b), returns the highest occurring digit and is expected to assume a low value under regular operation ("normal" condition); typically, the digit 1 appears most often in distributions of natural processes, in this case normal operation. A nonlinearity indicator (#21), taken from Schreiber and Schmitz (1997), based on a higher order autocovariance is calculated using Eq. (5)

$$t^{C3}(\tau) = \frac{1}{n-2\tau}\sum_{i=1}^{n-2\tau} x_{i+2\tau}\cdot x_{i+\tau}\cdot x_i \tag{5}$$

where $\boldsymbol{x}$ is the signal time series with n samples and τ is the time lag. A boolean indicator of symmetry (#23) in the signal is obtained with Eq. (6)

$$Bool_{sym} = \begin{cases} 1, if\ |mean(\boldsymbol{x}) - median(\boldsymbol{x})| \\ \quad < r|max(\boldsymbol{x}) - min(\boldsymbol{x})| \\ 0, otherwise \end{cases} \tag{6}$$

where $\boldsymbol{x}$ is the time series of measurement and r is a threshold, here considered equal to 0.01. The time-reversal asymmetry (#24), introduced by Fulcher and Jones (2014) is calculated using Eq. (7)

$$t_{rev}(\tau) = \frac{1}{n-2\tau}\sum_{i=1}^{n-2\tau} {x_{i+2\tau}}^2\cdot x_{i+\tau} - x_{i+\tau}\cdot x_i^2 \tag{7}$$

where $\boldsymbol{x}$ is the signal time series with n samples and τ is the time lag. Within the scope of this paper, the time lag is chosen as the minimum between the duration of the time-series segments divided by 100 and 3000 times the sampling period. This choice is guided by engineering heuristics and may be considered a "hyperparameter" that could be tuned for other applications. Still, as shown in

Section 4, these values apply to a large range of measurement applications.

Descriptive statistics of derived signals characterize signals, such as squared signals or derivatives, i.e., time-series that are numerically extracted form a reference measured signal. Thus, these statistical features, such as ean value of first (#29) and second (#30) derivative, are not applied to the raw time-series directly, but on a derived (computed) time series. Further features are extracted from moving averages that are calculated over short, intermediate and long time windows within the entire data series. In addition to standard-deviation of such moving averages (#34 - 36), the homogeneity (#37 - 39) is assessed using Eq. (8)

$$H_y = \frac{\max(y)}{mean(y)} \tag{8}$$

where $\boldsymbol{y}$ is the time-series of derived quantities. While moving averages of reduced length are more volatile, they should not deviate extensively from the mean of the entire time series, otherwise drifting or other sensor malfunctioning may be present. Within the scope of this paper, the short-term is set to 1s, the intermediate to 20 s and the long-term windows length is set to 120 s.

Descriptive statistics of frequency spectrum of the time signal characterize the signal in frequency domain. Frequency spectra are derived using the Fast-Fourier Transform (FFT) of the entire signal or via the Power-Spectrum Density (PSD) function, following Welch's averaged spectrum method (Welch 1967). While similar, the PSD averages the results over time windows after applying a window function, in this case the hamming function. For both spectra, FFT and PSD, the Kurtosis (#42 and #43) is derived and is used to indicate how the energy spectrum values are distributed, as some abnormal sensor behavior results in very distinctive peaks and long tails in frequency domain.

Heuristic characterization of the spectra provide insights into how the energy is distributed in frequency domain. Malfunctioning sensors are often characterized by a distinctive peak in frequency domain and thus, the entropy of FFT and PSD spectra (#40 and #41) and the homogeneity (see Eq. (8)) of both spectra (#44 and #45) are used to indicate how well the energy is spread throughout the frequency domain. In addition, the length of the frequency range (relative to the range from 0 Hz to the Nyquist frequency) containing the central 95% and 50% of the energy are calculated to characterize the energy spread (#46 – 49).

All 49 features described in Table 1 are used for training the OCC and the XGBoost classifier. In order to enable the distinctive interpretation of different anomalies that are initially unknown, preserving an extensive feature set is eminent and thus no feature reduction/selection is conducted.

4. Case studies

In this section we present the implementation of the proposed framework in data from two permanently monitored cable-stayed bridges in China. Schematic visualizations of the studied structures, including the approximate position of the sensors, are illustrated in Fig. 2. The data were provided for a blind competition along the 1st International Competition for Structural Health Monitoring (Bao *et al*. 2021). Two intrinsically different datasets are examined, namely acceleration recordings at positions showed in Fig. 2 (up) and force measurements of selected stay cables, marked in Fig. 2 (down). The labels provided by the organizers of the blind competition were considered as ground truth, while further verification of their validity was not deemed necessary. While vibration recordings reveal changes in the global dynamic response of the structure, the acting tension in stay cables is a valuable metric of their local structural health. Since the data was acquired during operation, the measured tension is affected by a number of factors, such as environmental effects and live loads. In case of damaged cables, redistribution of tension forces introduce further uncertainties that prevent the direct usage of force recordings as damage indexes.

Initially, the training and test sets for each dataset are presented. The OCC is trained on healthy data and an interpretable XGBoost model is trained to fit the class predictions offered by the OCC model. Considering the healthy class as reference, the SHAP scores of the test set are computed, as described in section 2. The proposed framework is applied to both datasets, for which labels describing the health state of the sensors are available.

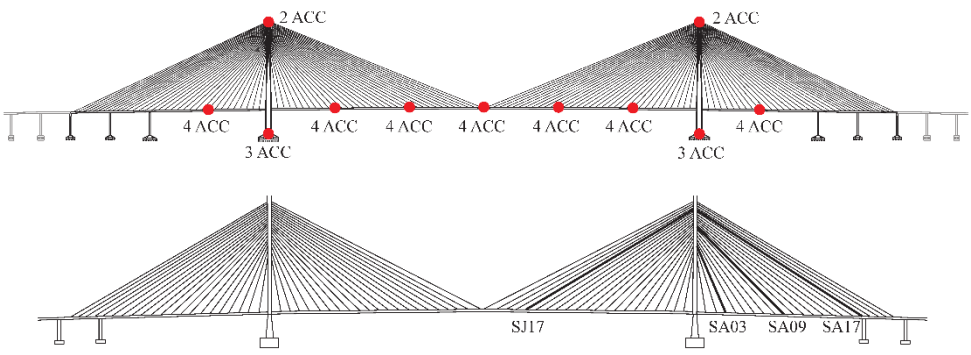


Fig. 2 Coarse illustration of the studied cable-stayed bridges. (up) The approximate positions of the accelerometers are marked in red color and (down) the monitored cables are highlighted with bold lines. Adapted from (Bao *et al*. 2021)

Within the scope of this paper, the labels are used exclusively for the validation of model predictions and explanations.

The evaluation of new data includes the following steps: truncation into segments of predefined length, transformation of time series data into the feature space, classification and the generation of decision trajectories. Given that the classification and interpretation with the XGBoost model require negligible time, the evaluation time is governed by the calculation of data features. Considering the vast feature set described in Section 3, the evaluation of new datasets for both studied cases took less than 30 seconds in a conventional desktop machine (Intel Xeon CPU E3-1275 v.5 @3.6 GHz), allowing for a near-real-time evaluation. It is mentioned that the sampling rate, the duration of the dataseries and potential extension of the feature space can affect the time required for the calculation of the features. The results demonstrate satisfactory predictive performance and the decision trajectories show impressive consistency in cases of samples under the same label. Additionally, the algorithm is shown to be robust when exposed to anomalies of different nature, such as damaged cables and sensor faults.

4.1 Acceleration data

This dataset comprises acceleration recordings of 38 channels, measuring at a sampling frequency of 20 Hz over a period of two months (January and February 2012). The approximate positions of the sensors are shown in Fig. 2. The data is structured in segments of 1 hour, which are labeled as "healthy" or "faulty", including further classification to common fault modes. Fig. 3 includes single examples of normal and faulty cases, based on the provided labels. It is mentioned that in many cases it is impossible to distinguish visually between "Trend" and "Drift" classes or "Minor" and "Outliers" classes, as shown in Fig. 4. For the training of the OCC only the "healthy" data from January are considered. All 38 channels are mixed and the data is segmented into 5-minute blocks, yielding 162900 time-series samples for training. The selected features (section 3) are computed and the trained OCC yields thresholds corresponding to the 90th and the 99th percentile of the "healthy" distribution. Data from February is considered unknown and is evaluated by the OCC without segmentation, after converting them into the feature space. Concatenating all 38 measuring channels yields 26448 data series of length equal to 3600. In order to train the XGBoost classifier to fit the OCC predictions, 80% of this data has been used for training and 20% for testing the performance of the XGBoost classifier. Table 2 includes the absolute numbers of the data used, as well as the fit metrics, after merging the "uncertain" classifications with "abnormal". The classifier exhibits satisfactory fit to the OCC, yielding an overall accuracy over 99%. By comparing the predictions with the ground truth for the same test set, the

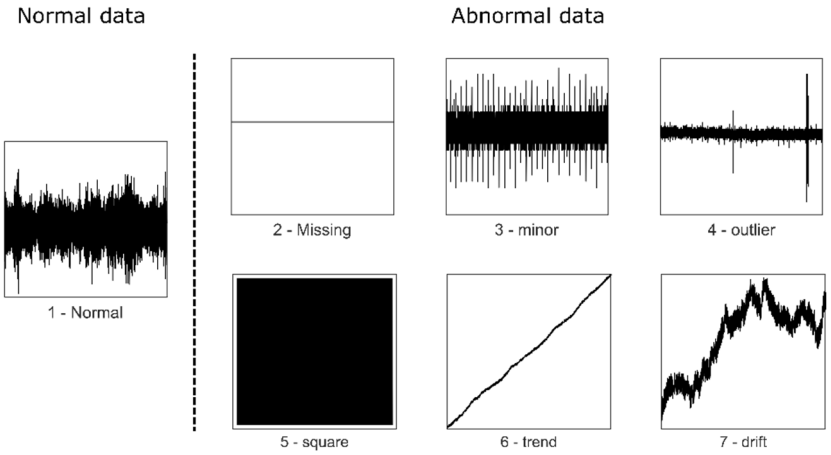
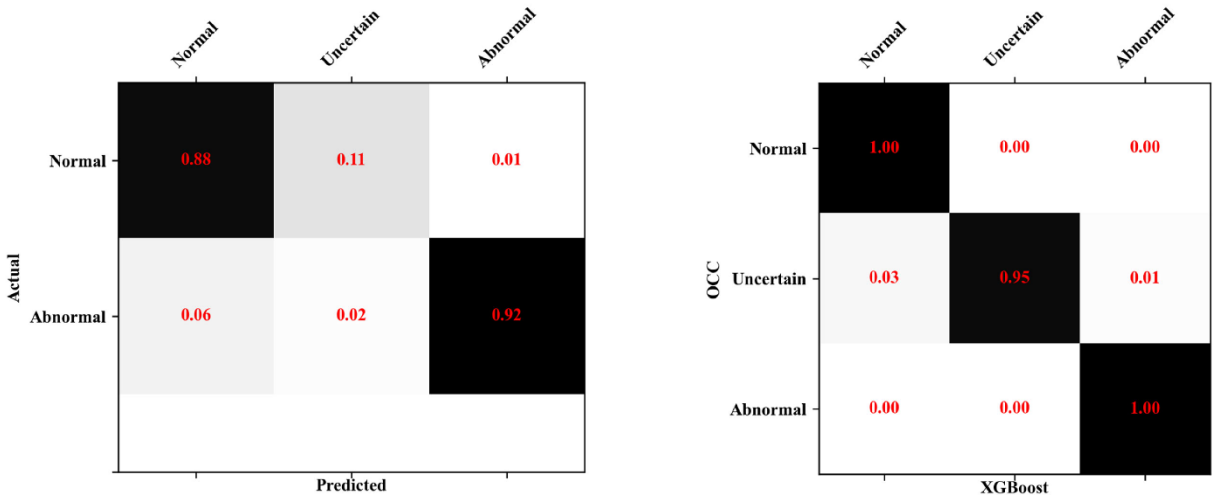


Fig. 3 Normal data and labeled anomalies observed in the provided acceleration data

(a) Data labelled as "Drift" (b) Data labelled as "Trend"

Fig. 4 Selected data readings from a single sensor at different times. No clear distinctive characteristics between the two fault classes are observed.



(a) OCC fit to test data. The provided labels are either "normal" or "abnormal". The "uncertain" class refers to thresholds defined in the output of OCC

(b) XGBoost fit to OCC predictions

Fig. 5 Acceleration dataset: predictive performance of the OCC and the XGBoost surrogate model

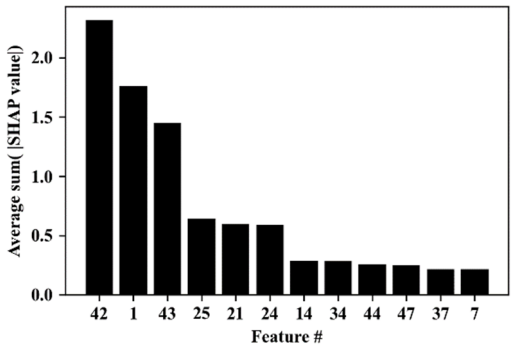


Fig. 6 Acceleration dataset: global feature importance. The 12 highest contributing features are plotted

results are summarized in the confusion matrix of Fig. 5(a). It can be observed that 88% of the "normal" data are properly identified, while 11% lies in the "uncertain" zone. The amount of False alarms is limited; 6% of the "abnormal" cases are mistakenly predicted as "normal". Finally, 92% of the "abnormal" samples are properly classified.

Beyond these satisfactory performance metrics, the proposed framework allows for further insights into the missclassifications. An interpretable XGBoost model is trained to fit the predictions of the OCC and the resulting fit is summarized in the confusion matrix of Fig. 5(b), yielding an overall accuracy over 99%. Considering the "healthy" predictions as baseline, the SHAP score is computed. After calculating the global feature importance with equation 3, the features are ordered in descending importance order (Fig. 6). The features that drive the model predictions are the Kurtosis in frequency domain (#42 and #43) the statistical mean (#1), the RMS amplitude (#25), the complexity metric c3 (#21) and the time reversal asymmetry statistic (#24).

Fig. 7 illustrates the decision trajectories of OCC predictions for the case of proper and false predictions, following the structure of the corresponding confusion matrix (Fig. 5(a)). The decision plots illustrate the impact of each feature on the accumulated SHAP score, which is associated with the model decisions. The features are ordered with decreasing importance, considering the "normal" class as the baseline. While the "normal" and "abnormal" samples show clearly different paths in the feature space, the impact of individual features in the "uncertain" classes could support understanding the model ambiguity. For the case of uncertain predictions of samples that are labeled as "normal", we observe that features #1, #42 and #43 are the main contributors that mislead the model predictions. For the case of false alarms, features #21, #24 and #25 are the ones that push the trajectory away from the "normal". Finally, regarding the missing alarms, the trajectories look rather similar to the normal trajectories, addressing the incapability of the model to observe these abnormalities, within the given feature space.

Subsequently, we study the decision paths of the specific fault types. The labels of the faults are available and allow plotting together the trajectories of all samples that belong to the same class (Fig. 8). Comparing the decision trajectories of the faults with the "healthy" case, we observe that all fault classes point to negative SHAP values, whereas the healthy samples point towards positive SHAP values. This finding confirms that the model is capable of separating "normal" from "abnormal" behavior for all fault classes. Studying the trajectories closer allows us to detect differences between fault classes, which indicate that the impact of certain features can reveal information about the type of the fault.

The decision trajectory of "Square" faults shows significant variability regarding the impact of features #21 and #24 (nonlinearity metrics) compared to the previous trajectories. A more consistent difference between "Square" and "Trend/Drift" paths can be spotted at the impact of

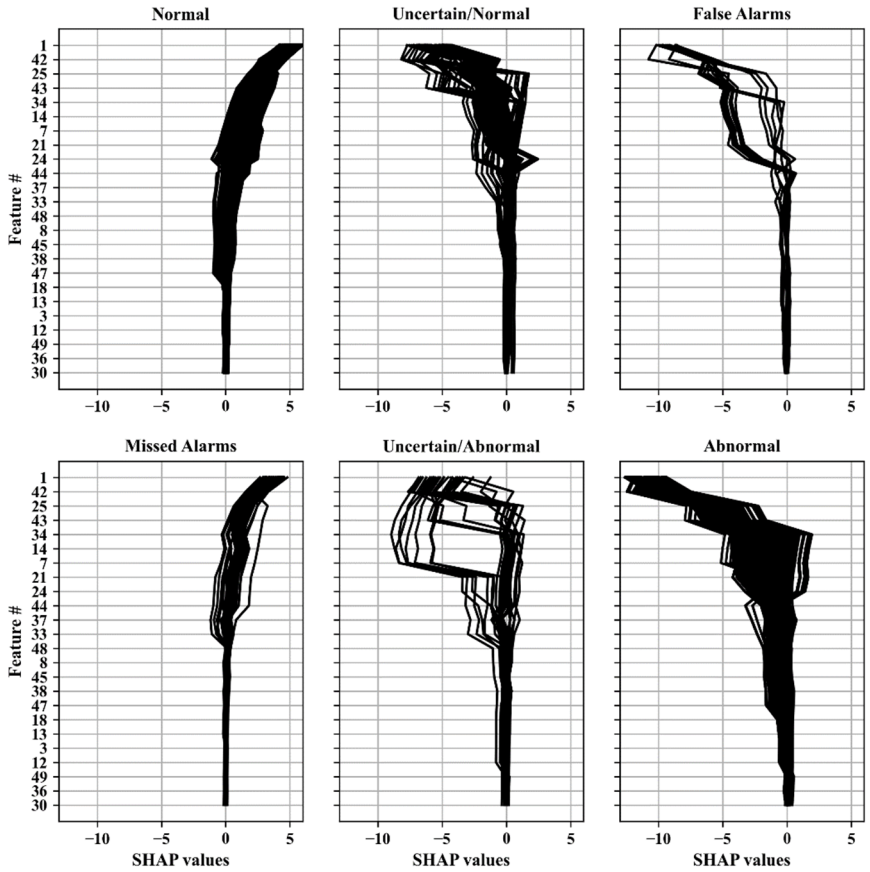


Fig. 7 Acceleration dataset: decision trajectories of OCC predictions. The 24 highest contributing features are plotted

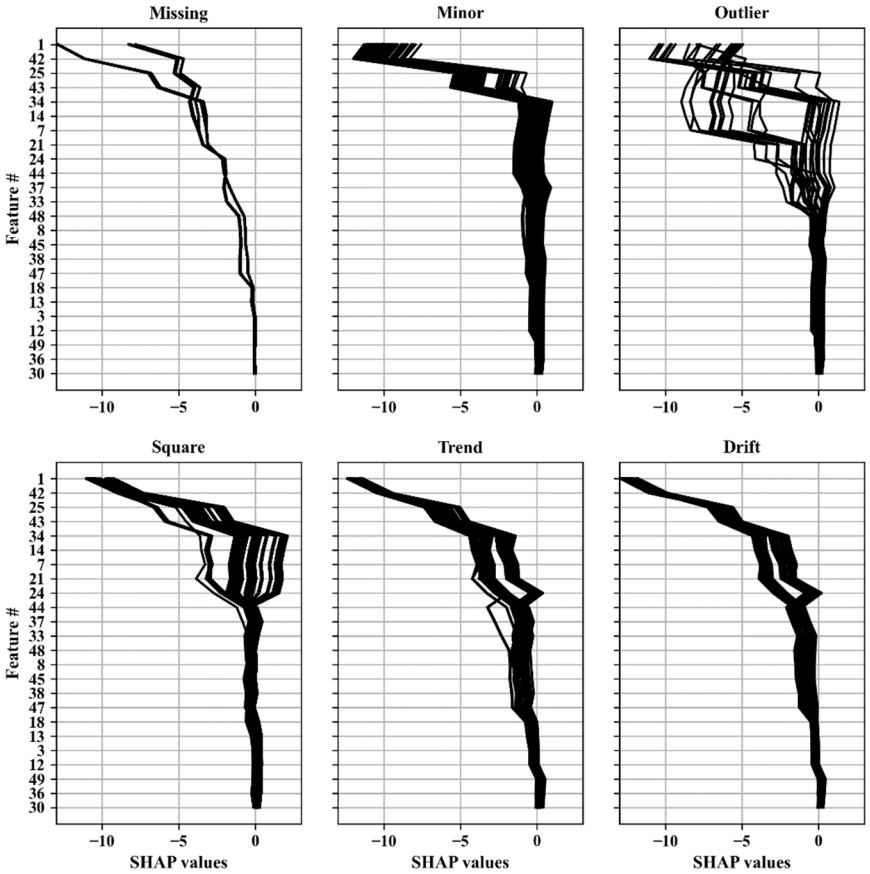
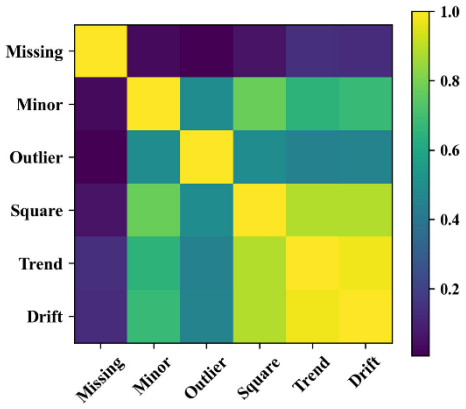
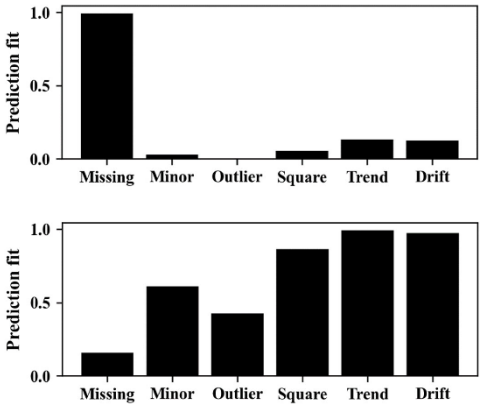


Fig. 8 Acceleration dataset: decision trajectories of labeled fault classes. The 24 highest contributing features are plotted



(a) Similarity control of the reference fault trajectories through DTAC



(b) Characteristic fault predictions based on DTAC. The real labels correspond to "Missing" and "Trend" classes respectively

Fig. 9 Acceleration dataset: predictive performance by applying the DTAC criterion

feature 12, which refers to the length of the longest strike below mean. The "Minor" and "Outlier" trajectories share a distinctive behavior with respect to the impact of features #1 and #25, referring to the statistical mean and RMS amplitude respectively. The accumulated SHAP score of "Outliers" shows a relatively larger scatter when compared to "Minor", which can be attributed to the impact of feature #7 (Kurtosis in time domain) and feature #33, which relates to the amount of outliers. Finally, for the case of "Missing", two different paths are observed. The first one assimilates very much to the "Trend/Drift" path, showing, however, a distinctive behavior with respect to feature #33. The second path shows a characteristic zigzag regarding the impact of the features #1, #42, #25, #43 that is opposite to the "Minor" case and different from the "Trend/Drift" faults.

The above observations confirm the existence of consistent patterns in the feature attributions for different fault types. By considering the median values of the feature attributions for each fault class, the reference trajectories are extracted. Fig. 9(a) illustrates the DTAC matrix computed on the first derivative of the reference trajectories. The DTAC exposes that "Trend/Drift" are highly correlated, while the class "Square" is also difficult to distinguish, within the given feature space. Fig. 9(b) demonstrates two characteristic fault predictions, based on the fit between individual trajectories (classified as "Missing" and "Trend" respectively) and the reference ones. While for the case of "Missing" the prediction is almost certain, for the case classified as "Trend" both "Square" and "Drift" show high predictive fit. Although this ambiguity exposes the classification limits of the current model, it highlights the importance of explanations in this regard. Further data and targeted enrichment of the feature space could sharpen the classification performance and allow for additional gains from OCC predictions.

Overall, it is demonstrated that the decision trajectories expose distinctive characteristics of different fault types, even when no information regarding the faults is provided during the training of the underlying OCC. While the distinction between assimilating faults, such as "Trend" and "Drift", is not possible, the decision trajectories spot unique characteristics between the rest of the classes, enabling a preliminary interpretation of faults in the absence of labeled data. Visualizing the trajectory paths provides a valuable tool for the enrichment of the binary classification offered by conventional OCC.

4.2 Cable force data

Stay cables are critical components, suffering from the coupled effects of fatigue and corrosion along with harsh environmental conditions. Permanent condition monitoring of stay cables offers an important complement to on-site inspections, in order to ensure the long-term safety and functionality of sustained bridges. This dataset includes recordings of load-cells placed at the anchorages of 14 selected cables of a stay-cable bridge in China (Fig. 2). The available data was recorded at a sampling frequency of 2 Hz and covers 10 days between 2006 and 2011 (2006-05-13 to 2006-05-19, 2007-12-14, 2009-05-05, and 2011-11-01). Fig. 10 shows the characteristic response of a "healthy" cable measured in 2006. The observed variability is attributed to varying EOCs. All available data from 2006 are labeled as healthy and are used for the training of the OCC. The data of all sensors are split into 100-second segments and concatenated, yielding 60480 data-series for training. These data segments are subsequently converted into the feature space described in section 3, and then used for the training of the OCC. The thresholds corresponding to the 90th and the 99th percentile of the training distribution are computed and defined as lower bounds for the

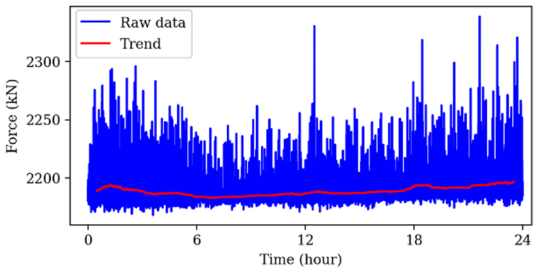


Fig. 10 Cable force dataset: characteristic "healthy" cable tension response measured on 2006-05-15

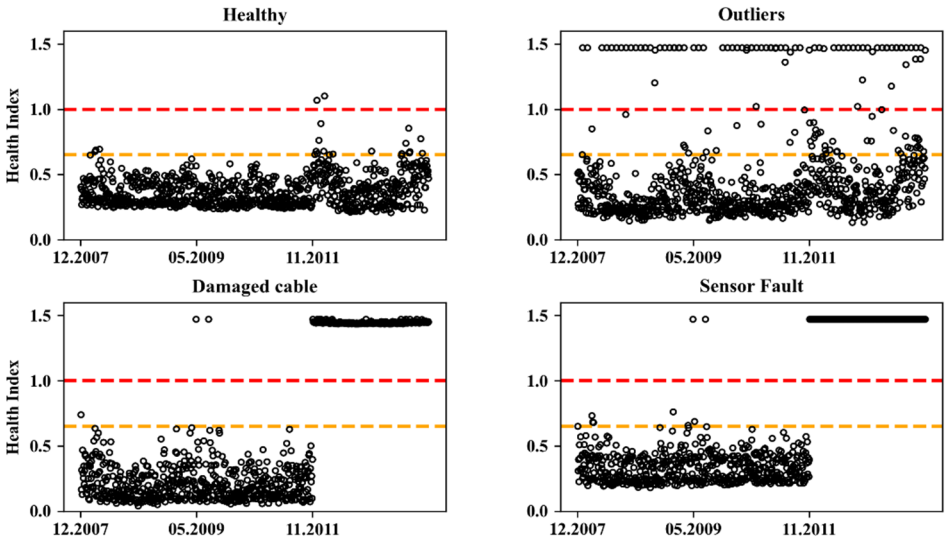


Fig. 11 Cable force dataset: Evolution of health index of four characteristic cases, labeled as "healthy", "outliers", "damaged cable" and "sensor fault'. The orange and red dashed lines correspond to the 90th and the 99th percentile of the training distribution and define the limits of the "uncertain" and "faulty" classes respectively

Table 2 Summary of the data used for training and testing the XGBoost classifier. The false positives/ negatives refer to the fit between XGBoost and OCC

Dataset	Training data	Test data	Duration [sec]	False positives	False negatives
Acceleration	21158	5290	3600	13	4
Cable force	9676	2420	100	8	24

"uncertain" and "faulty" classes respectively.

The data recorded from 2007 to 2011 are used for testing the algorithm. The anomaly detection score of the OCC for representative channels is plotted in Fig. 11. According to the available labels based on a visual inspection conducted in 2011, one sensor persistently included a significant amount of outliers, two sensors were reported as failing and one monitored cable was classified as damaged. An interpretable XGBoost model is trained to fit the predictions of the OCC. Considering all 14 measured cables yields 12096 data series of length equal to 100 seconds. From this data, 80% is used for training and 20% for testing the performance of the XGBoost classifier. Table 2 includes the description of the data used, as well as the fit metrics, after merging the "uncertain" classifications with "abnormal". The classifier exhibits satisfactory fit to the OCC, yielding an overall accuracy over 99%. The normalized fit of the XGBoost classifier to the OCC is summarized in the confusion matrix shown in Fig. 12.

Considering the reference data as baseline, the Shapley values of the test set are computed. Initially, the global feature importance is estimated, as explained in section 2. Fig. 14(a) reports the features in descending importance order, with the most important being the statistical most common value (#3), the standard deviation (#4) and the coefficient of variation in time domain (#8), as well as the features #32, #36 and #48 (per Table 1).

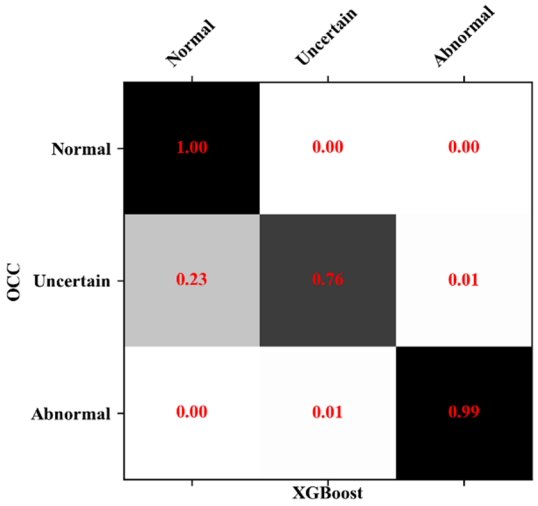


Fig. 12 Cable force dataset: XGBoost fit to OCC predictions

Fig. 13 demonstrates the decision trajectories of "healthy" and "faulty" classified samples. All faulty cases result in negative SHAP values, whereas the "healthy" samples point towards positive SHAP values. The histogram of the accumulated SHAP score for all labeled classes is given in Fig. 14(b), illustrating the capability of the model to detect abnormal behavior. Zooming closer into the decision trajectories, differences between fault classes appear and indicate that the impact and the dependencies of certain features can reveal information about the type of the fault. The decision trajectories of "outliers" are characterized by the impact of features #4, #8 and #48, which correspond to the standard deviation, the coefficient of variation and the frequency range including 95% of the signal energy (see Table 1). The trajectory-footprint of the samples corresponding to "damaged cable" is significantly thicker compared to other cases, indicating larger variability, which is expected, as structural damage

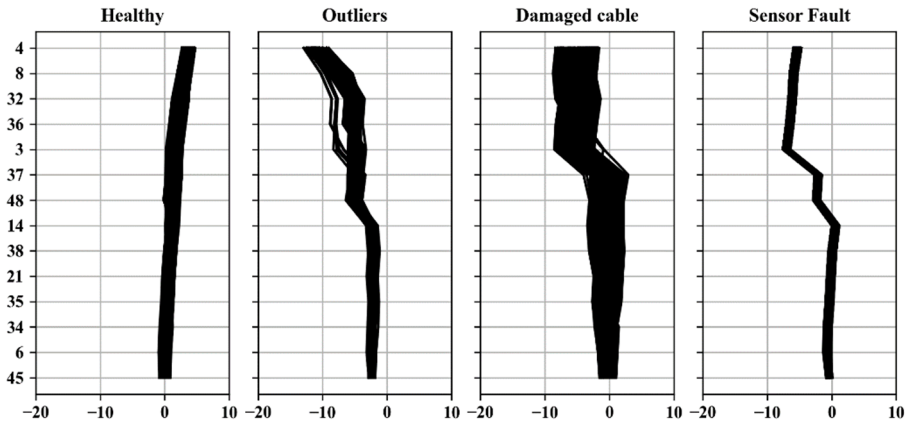
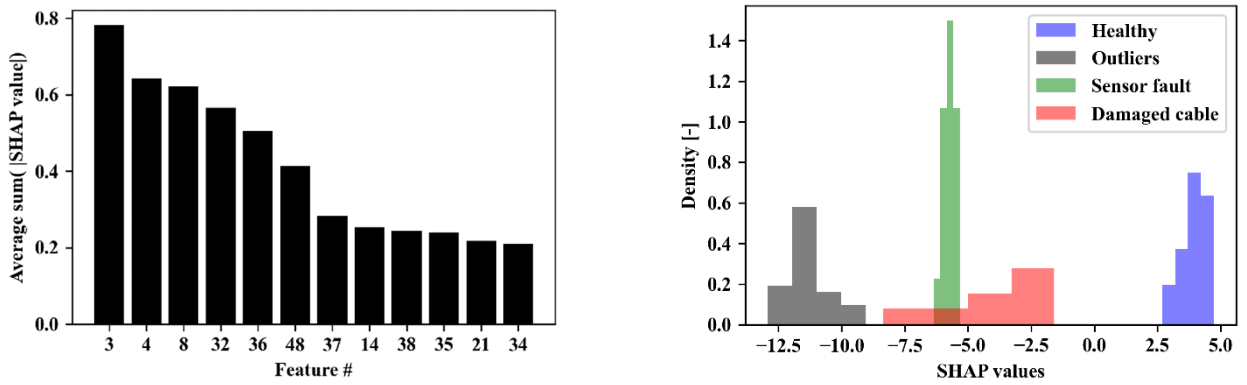


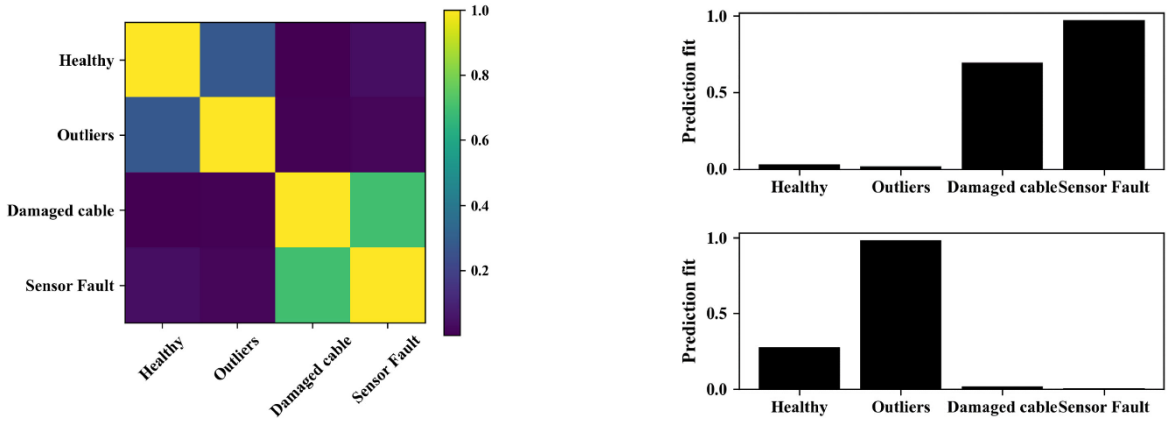
Fig. 13 Cable force dataset: decision trajectories of labeled fault classes. The 14 highest contributing features are plotted.



(a) Global feature importance

(b) Histogram of the accumulated SHAP score for all labeled classes

Fig. 14 Cable force dataset: global feature importance and accumulated SHAP score



(a) Similarity control of the reference fault trajectories through DTAC

(b) Characteristic fault predictions based on DTAC. The real labels correspond to “sensor fault” and “outliers” classes respectively

Fig. 15 Cable force dataset: predictive performance by applying the DTAC criterion

introduces spurious novelties into the response. The impact of feature #3 (statistical most common value) seems to provide a consistent bias on the model decision, indicating the sensor invariance to ambient and traffic loading, as a result of the cable relaxation. For the case of sensor fault, no further information regarding the fault type is provided. The decision trajectory is rather narrow, showing low variability, and is driven by the impact of features #3 (statistical most common value) and #48 (frequency range of high energy response). This could be attributed to the invariance of sensor readings to loading and to the imposed noise due to the sensor failure.

The above observations confirm the existence of consistent patterns in the feature attributions for different

fault types. By considering the median values of the feature attributions for each fault class, the reference trajectories are extracted. The DTAC matrix, computed on the first derivative of the reference trajectories, illustrates that "damaged cable" and "sensor fault" are partially correlated, while the rest of classes seem to be adequately separated (Fig. 15(a)). Fig. 15(b) demonstrates two characteristic fault prediction based on the fit between individual trajectories (classified as "sensor faults" and "outliers" respectively) and the reference ones. While for both cases the right prediction is associated with the maximum fit, in the case of "sensor faults" the "damaged cable" class reaches high values, which is attributed to the similarity of the corresponding reference trajectories.

Overall, the application to the dataset of load-cell sensors demonstrated good performance in capturing anomalies related to various sources, while the interpretability framework exposed the feature dependencies that drive the model predictions. Although the proposed framework aims at identifying sensor faults, the algorithm also proves capable of detecting structural damage, when the latter is observable within the predefined feature space. Comparing the decision trajectories of different fault sources highlighted the potential of the proposed framework in enriching the binary classification of OCC with information that supports understanding the nature of the fault.

5. Conclusions

This work introduces a framework for practical sensor-fault detection and interpretation within the context of SHM. Based on the well-established SVM method for anomaly detection, a coalitional game theory approach is applied to the interpretation of the model decisions, thereby providing insights into to the individual feature impact and feature dependencies on the classification of new samples. A comprehensive set of features, suitable for generic time series data, is assembled, in order to convert signals into an informative feature space prior to the model evaluation. The proposed framework is deployed on two intrinsically different datasets, comprising acceleration recordings and stay-cable force measurements from two long-term monitored bridges. Good accuracy in terms of anomaly detection is achieved in both cases, reflecting the seamless adaptation of the algorithm to various data domains. The sum of the feature contributions provides a straightforward estimation of the feature importance for individual predictions, which is a significant contribution towards explaining machine-learning decisions, compared with conventional permutation methods that estimate the feature importance globally.

For the first time, the term "decision trajectories" is introduced in the context of SHM, illustrating the feature impact and the feature dependencies that drive individual model predictions. In order to evaluate the consistency between different decision trajectories, a measure of correlation is proposed, namely the Decision Trajectory Assurance Criterion (DTAC), which enables automatizing fault classification, provided that reference trajectories of characteristic faults are available. This functional enhancement of one-class classifiers can be of significant importance for the interpretation of anomalies in an application-agnostic manner. Thus, limitations of conventional feature heuristics, which require prior knowledge of the expected faults, may be overcome.

In a nutshell, given a set of data classified as "normal", the proposed frameworks enables a seamless detection of faults in new data. By analyzing the decision trajectories, monitoring experts may translate the insights offered by individual feature impacts and dependencies into a better understanding of the underlying data structure. Fusing expert judgment with the information obtained from the decision trajectories may enable rapid labeling of characteristic faults or false alarms that could be further utilized for the supervised training of more advanced multi-class classifiers or for other data-mining purposes.

The proposed framework provides an easy-to-train and application-agnostic anomaly detector that can be integrated into the preprocessing part of various SHM and condition monitoring applications, providing a first screening of the sensor health, prior to further analysis. The expected sensor faults must be observable within the predefined feature space. While a vast feature space is provided in this paper, the proposed framework is independent from the data features and the features space can be seamlessly extended or adjusted to project-specific demands. The algorithm is expected to classify as faulty any dataset deviating from the expected reference response. Apart from sensor faults, anomalies could be detected due to extreme EOCs or structural damage. Evaluating the corresponding decision trajectories, as well as EOCs, monitoring data could facilitate the understanding of the nature of new faults. Finally, reference data that cover the healthy response of the structure under various EOCs are required. In order to ensure that in real applications, expert judgement could be deployed. While the proposed framework supports the consideration of time varying effects (such as EOCs) by simply retraining the anomaly detector as soon as further "normal" data are available, the adoption of domain adaptation or reinforced learning approaches could potentially allow for transfer-learning between associated monitoring projects. Exploring this potential, although falling outside the scope of the presented work will be the focus of future research endeavors.

Acknowledgments

The authors would like to thank the organizations of the International Project Competition for SHM (IPC-SHM 2020) ANCRiSST, Harbin Institute of Technology (China), and University of Illinois at Urbana-Champaign (USA) for their generously providing the invaluable data from actual structures. The authors also would like to thank the chairs of IPC-SHM 2020 Prof. Hui Li and Prof. Billie F. Spencer Jr for their leadership on the competition.

The work presented in this paper was financially supported by the Real-time Earthquake Risk Reduction for a Resilient Europe 'RISE' project, financed under the European Union Horizon 2020 research and innovation

program, under grant agreement No 821115, as well as the ETH Risk Center project 'DynaRisk', financed under grant agreement ETH-11 18-1.

References

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M. and Inman, D.J. (2017), "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks", *J. Sound Vib.*, **388**, 154-170. http://dx.doi.org/10.1016/j.jsv.2016.10.043

Allemang, R.J. (1982), "A correlation coefficient for modal vector analysis", *Proceedings of the 1st International Modal Analysis Conference*.

Alonso, E.E., Pinyol, N.M. and Puzrin, A.M. (2010), *Geomechanics of Failures. Advanced Topics*, Dordrecht, Springer Netherlands. http://link.springer.com/10.1007/978-90-481-3538-7

An, Y., Chatzi, E., Sim, S.H., Laflamme, S., Blachowski, B. and Ou, J. (2019), "Recent progress and future trends on damage identification methods for bridge structures", *Struct. Control Health Monitor.*, **26**(10), 1-30. https://doi.org/10.1002/stc.2416

Athanasiou, M., Sfrintzeri, K., Zarkogianni, K., Thanopoulou, A.C. and Nikita, K.S. (2020), "An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus", *Proceedings of 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering* (*BIBE*), Cincinnati, OH, USA, October, pp. 859-864. https://ieeexplore.ieee.org/document/9288053/

Avendaño-Valencia, L.D., Chatzi, E.N., Koo, K.Y. and Brownjohn, J.M. (2017), "Gaussian process time-series models for structures under operational variability", *Front. Built Environ.*, **3**. http://journal.frontiersin.org/article/10.3389/fbuil.2017.00069/full

Azimi, M., Eslamlou, A.D. and Pekcan, G. (2020), "Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review", *Sensors*, **20**(10), 2778. https://doi.org/10.3390/s20102778

Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z. and Li, H. (2019a), "The state of the art of data science and engineering in structural health monitoring", *Engineering*, **5**(2), 234-242. https://doi.org/10.1016/j.eng.2018.11.027

Bao, Y., Tang, Z., Li, H. and Zhang, Y. (2019b), "Computer vision and deep learning–based data anomaly detection method for structural health monitoring", *Struct. Health Monitor.*, **18**(2), 401-421. https://doi.org/10.1177/14759217211006485

Bao, Y., Li, J., Nagayama, T., Xu, Y., Spencer Jr, B.F. and Li, H. (2021), "The 1st International Project Competition for Structural Health Monitoring (IPC-SHM, 2020): A Summary and Benchmark Problem", *Struct. Health Monitor.*, **20**(4), 2229-2239. https://doi.org/10.1177/14759217211006485

Bull, L., Worden, K., Manson, G. and Dervilis, N. (2018), "Active learning for semi-supervised structural health monitoring", *J. Sound Vib.*, **437**, 373-388. https://doi.org/10.1016/j.jsv.2018.08.040

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2020), "Explainable AI in fintech risk management", *Front. Artif. Intell.*, **3**, 26. https://doi.org/10.3389/frai.2020.00026

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H. (2018), "Xgboost: extreme gradient boosting", *R. Package Version 0.71-2*, pp. 1-4.

Dietterich, T. (1995), "Overfitting and undercomputing in machine learning", *ACM Computing Surveys* (*CSUR*) **27**(3), 326-327.

Fan, W. and Qiao, P. (2011), "Vibration-based damage identification methods: a review and comparative study", *Struct. Health Monitor.*, **10**(1), 83-111. https://doi.org/10.1177/1475921710365419

Figueiredo, E. and Santos, A. (2018), "Machine learning algorithms for damage detection", In: *Vibration-Based Techniques for Damage Detection and Localization in Engineering Structures*, pp. 1-39. https://doi.org/10.1142/9781786344977_0001

Figueiredo, E., Park, G., Farrar, C.R., Worden, K. and Figueiras, J. (2011), "Machine learning algorithms for damage detection under operational and environmental variability", *Struct. Health Monitor.*, **10**(6), 559-572. https://doi.org/10.1177/1475921710388971

Fulcher, B.D. and Jones, N.S. (2014), "Highly comparative feature-based time-series classification", *IEEE Transact. Knowl. Data Eng.*, **26**(12), 3026-3037. https://doi.org/10.1109/TKDE.2014.2316504

Goldberger, J., Hinton, G.E., Roweis, S. and Salakhutdinov, R.R. (2005), "Neighbourhood components analysis", In: *Advances in Neural Information Processing Systems*, MIT Press.

Gui, G., Pan, H., Lin, Z., Li, Y. and Yuan, Z. (2017), "Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection", *KSCE J. Civil Eng.*, **21**(2), 523-534. https://doi.org/10.1007/s12205-017-1518-5

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B. (1998), "Support vector machines", *IEEE Intell. Syst. their Applicat.*, **13**(4), 18-28. https://doi.org/10.1109/5254.708428

Hill, T.P. (1995a), "A statistical derivation of the significant-digit law", *Statist. Sci.*, **10**(4), 354-363. https://doi.org/10.1214/ss/1177009869

Hill, T.P. (1995b), "The significant-digit phenomenon", *Am. Mathe. Monthly*, **10**(4), 322-327. https://doi.org/10.1080/00029890.1995.11990578

Jaishi, B. and Ren, W.X. (2006), "Damage detection by finite element model updating using modal flexibility residual", *J. Sound Vib.*, **290**(1-2), 369-387. https://doi.org/10.1016/j.jsv.2005.04.006

Lim, S. and Chi, S. (2019), "Xgboost application on bridge management systems for proactive damage estimation", *Adv. Eng. Inform.*, **41**, 100922. https://doi.org/10.1016/j.aei.2019.100922

Limongelli, M.P., Chatzi, E., Döhler, M., Lombaert, G. and Reynders, E. (2016), "Towards extraction of vibration-based damage indicators", *Proceedings of the 8th European Workshop on Structural Health Monitoring*, *EWSHM 2016*, Bilbao, Spain, July, Vol. 1, pp. 546-555.

Long, J. and Buyukozturk, O. (2014), "Automated structural damage detection using one-class machine learning", *Proceedings of the Society for Experimental Mechanics Series*, Vol. 4, pp. 117-128. https://doi.org/10.1007/978-3-319-04546-7_14

Lundberg, S.M. and Lee, S.I. (2017), "A unified approach to interpreting model predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 4768-4777.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J. and Lee, S.I. (2018), "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery", *Nature Biomed. Eng.*, **2**(10), 749-760. https://doi.org/10.1038/s41551-018-0304-0

Martakis, P., Reuland, Y. and Chatzi, E. (2021), "Amplitude-dependent model updating of masonry buildings undergoing demolition", *Smart Struct. Syst., Int. J.*, **27**, 157-172. https://doi.org/10.12989/sss.2021.27.2.157

Moaveni, B., Conte, J.P. and Hemez, F.M. (2009), "Uncertainty and sensitivity analysis of damage identification results

obtained using finite element model updating", *Comput.-Aided Civil Infrastr. Eng.*, **24**(5), 320-334. https://doi.org/10.1111/j.1467-8667.2008.00589.x

Movsessian, A., Cava, D.G., Tcherniak, D. and Janeliukstis, R. (2020), "A methodology on interpretable novelty detection", *Proceedings of the 1th International Conference on Structural Dynamics*, Athens, Greece, November, pp. 922-935. https://doi.org/10.47964/1120.9073.19621

Movsessian, A., Cava, D.G. and Tcherniak, D. (2021), "Interpretable machine learning in damage detection using Shapley Additive Explanations", *Preprint*. https://doi.org/10.31224/osf.io/96yf5

Musavi, M.T., Ahmed, W., Chan, K.H., Faris, K.B. and Hummels, D.M. (1992), "On the training of radial basis function classifiers", *Neural Networks*, **5**(4), 595-603. https://doi.org/10.1016/S0893-6080(05)80038-3

Neves, A.C., Gonzalez, I., Leander, J. and Karoumi, R. (2017), "Structural health monitoring of bridges: a model-free ANN-based approach to damage detection", *J. Civil Struct. Health Monitor.*, **7**(5), 689-702. https://doi.org/10.1007/s13349-017-0252-5

Onchis, D.M. and Gillich, G.R. (2021), "Stable and explainable deep learning damage prediction for prismatic cantilever steel beam", *Comput. Ind.*, **125**, 103359. https://doi.org/10.1016/j.compind.2020.103359

Pai, S.G., Reuland, Y. and Smith, I.F. (2019), "Data-interpretation methodologies for practical asset-management", *J. Sensor Actuator Networks*, **8**(2), 1-29. https://doi.org/10.3390/jsan8020036

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A.K. (2020), "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis", *Accid. Anal. Preven.*, **136**, 105405. https://doi.org/10.1016/j.aap.2019.105405

Reuland, Y., Lestuzzi, P. and Smith, I.F. (2017), "Data-interpretation methodologies for non-linear earthquake response predictions of damaged structures", *Front. Built Environ.*, **3**, 43. https://doi.org/10.3389/fbuil.2017.00043

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), ""Why should i trust you?" Explaining the predictions of any classifier", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778

Roth, A.E. (1988), *The Shapley Value*: *Essays in Honor of Lloyd*, S. Shapley. Ed., Cambridge University Press.

Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J. and Platt, J.C. (2000), "Support vector method for novelty detection", *Adv. Neural Inform. Process. Syst.*, Vol. 12, pp. 582-588.

Schreiber, T. and Schmitz, A. (1997), "Discrimination power of measures for nonlinearity in a time series", *Phys. Rev. E*, **55**(5), 5443-5447. https://doi.org/10.1103/PhysRevE.55.5443

Shapley, L.S. (1953), "17. A value for n-person games", In: *Contributions to the Theory of Games* (*AM-28*), Volume II, Princeton University Press, pp. 307-318. https://doi.org/10.1515/9781400881970-018

Smith, I.F. (2016), "Studies of sensor data interpretation for asset management of the built environment", *Front. Built Environ.*, **2**, 1-9. https://doi.org/10.3389/fbuil.2016.00008

Sohn, H., Farrar, C.R., Hemez, F.M., Shunk, D.D., Stinemates, D.W., Nadler, B.R. and Czarnecki, J.J. (2001), "A review of structural health monitoring literature: 1996–2001", LA-UR-02-2095.

Štrumbelj, E. and Kononenko, I. (2014), "Explaining prediction models and individual predictions with feature contributions", *Knowl. Inform. Syst.*, **41**(3), 647-665. https://doi.org/10.1007/s10115-013-0679-x

Tibaduiza, D.A., Torres-Arredondo, M.A., Mujica, L.E., Rodellar, J. and Fritzen, C.P. (2013), "A study of two unsupervised data driven statistical methodologies for detecting and classifying damages in structural health monitoring", *Mech. Syst. Signal Process.*, **41**(1-2), 467-484. http://dx.doi.org/10.1016/j.ymssp.2013.05.020

Tibaduiza, D., Torres-Arredondo, M.Á., Vitola, J., Anaya, M. and Pozo, F. (2018), "A damage classification approach for structural health monitoring using machine learning", *Complexity*, 2018. https://doi.org/10.1155/2018/5081283

Vilone, G. and Longo, L. (2020), "Explainable artificial intelligence: a systematic review", arXiv preprint arXiv:2006.00093. http://arxiv.org/abs/2006.00093.

Welch, P. (1967), "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms", *IEEE Transact. Audio Electroacoust.*, **15**(2), 70-73. https://doi.org/10.1109/TAU.1967.1161901

Wilson, G.T. (2016), "Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1", *J. Time Series Anal.*, **37**(5), 709-711. https://doi.org/10.1111/jtsa.12194

Worden, K., Manson, G. and Fieller, N.R. (2000), "Damage detection using outlier analysis", *J. Sound Vib.*, **229**(3), 647-667. https://doi.org/10.1006/jsvi.1999.2514

Ying, Y., Garrett Jr, J.H., Oppenheim, I.J., Soibelman, L., Harley, J.B., Shi, J. and Jin, Y. (2013), "Toward data-driven structural health monitoring: application of machine learning and signal processing to damage detection", *J. Comput. Civil Eng.*, **27**(6), 667-680. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000258

Zgonnikov, A., Aleni, A., Piiroinen, P.T., O'Hora, D. and di Bernardo, M. (2017), "Decision landscapes: visualizing mouse-tracking data", *Royal Soc. Open Sci.*, **4**(11), 170482. https://doi.org/10.1098/rsos.170482

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B. and Si, Y. (2018), "A data-driven design for fault detection of wind turbines using random forests and XGboost", *IEEE Access*, **6**, 21020-21031. https://doi.org/10.1109/ACCESS.2018.2818678