

Interpretable machine learning in damage detection using Shapley Additive Explanations

Artur Movsessian

Institute for Infrastructure
and Environment

School of Engineering

University of Edinburgh

United Kingdom EH9 3FG

Email: Artur.Movsessian@ed.ac.uk

David García Cava

Institute for Infrastructure
and Environment

School of Engineering

University of Edinburgh

United Kingdom EH9 3FG

Email: david.garcia@ed.ac.uk

Dmitri Tcherniak

Hottinger Brüel & Kjær

Sound and Vibration Measurements

Skodsborgvej 307, Nærum 2850, Denmark

Email: dtcherniak@bksv.com

ABSTRACT

In recent years, Machine Learning (ML) techniques have gained popularity in Structural Health Monitoring (SHM). These have been particularly used for damage detection in a wide range of engineering applications such as wind turbine blades. The outcomes of previous research studies in this area have demonstrated the capabilities of ML for robust damage detection. However, the primary challenge facing ML in SHM is the lack of interpretability of the prediction models hindering the broader implementation of these techniques. For this purpose, this study integrates the novel Shapley Additive exPlanations (SHAP) method into a ML-based damage detection process as a tool for introducing interpretability and, thus, build evidence for reliable decision-making in SHM applications. The SHAP method is based on coalitional game theory and adds global and local interpretability to ML-based models by computing the marginal contribution of each feature. The

contribution is used to understand the nature of damage indices (DIs). The applicability of the SHAP method is first demonstrated on a simple lumped mass-spring-damper system with simulated temperature variabilities. Later, the SHAP method has been evaluated on data from an in-operation V27 wind turbine with artificially introduced damage in one of its blades. The results show the relationship between the environmental and operational variabilities (EOVs) and their direct influence on the damage indices. This ultimately helps to understand the difference between false positives caused by EOVs and true positives resulting from damage in the structure.

NOMENCLATURE

SHM	Structural health monitoring
O&M	Operation and maintenance
WT	Wind turbine
EOV	Environmental and operational variabilities
ML	Machine learning
CNNs	Convolutional Neural Networks
SHAP	Shapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
DSF	Damage sensitive features
MD	Mahalanobis distance
DI	Damage index
MSE	Mean squared error
XGBoost	eXtreme Gradient Boosting

1 INTRODUCTION

Structural health monitoring (SHM) refers to the process of identifying changes in the integrity of a structure based on observations and measurements that describe its current state. SHM has sparked the interest of researchers due to the untapped potential for safety and reliability im-

provements and operation and maintenance (O&M) cost reductions. Among the diversity of methods adopted by researchers, data-driven vibration-based structural health monitoring (VSHM) is a promising approach and is continuously being studied. The primary goal of VSHM is to detect damage in a structure through the following steps [1]: (1) Measure vibration responses from the structure; (2) Extract damage sensitive features (DSFs), i.e. metrics that contain useful information to characterize the state of the structure; (3) Select a method to process the DSFs and calculate a damage index (DI); and (4) compare the DI with a predefined threshold, thereby identifying outliers that indicate potential damage in the structure. For the last two steps, commonly, a statistical outlier analysis method is implemented such as the Mahalanobis distance (MD) to identify discordant outliers in a feature set [2]. However, one of the prevailing challenges in damage detection is the influence of environmental and operational variabilities (EOVs) on DSFs when monitoring structures and collecting vibration responses. Several studies have demonstrated how EOVs can camouflage the presence of damage and, ultimately, hinder damage detection [3–6].

To address this challenge, researchers have investigated the use of machine learning (ML) techniques to identify and mitigate the influence of EOVs. In the past decade, numerous studies have implemented these techniques showing promising results when monitoring structures, identifying failures or damage [7, 8]. A study by Santos et al. [9] compared several ML algorithms, namely one-class support vector machine, support vector data description, kernel principal component analysis and greedy kernel principal component analysis to detect damage in a three-story aluminum frame structure under the influence of EOVs. The results have shown reliability in recognizing patterns and clustering the observations from the healthy structure, thereby mitigating the effects of EOVs and ultimately increasing the damage detection accuracy. Abdeljaber et al. [10] presented a computationally inexpensive system using 1D Convolutional Neural Networks (CNNs) for real-time damage detection and localisation. The approach enables the automatic extraction of optimal DSFs and was tested on eight structural cases derived from large-scale experiments on a simulator. The results showed the high level of generalization of this approach and the advantage of eliminating the need for manual model or parameter tuning on feature extraction. In the field of wind energy, Zhang et al. [11] proposed a data-driven framework using random forests (RFs)

in combination with extreme gradient boosting (XGBoost) for fault detection in wind turbines. The former is used to rank the DSFs according to their relevance and, later, the top-ranking features are used to train the XGBoost algorithm. The approach was tested on numerical simulations of three different types of wind turbines (onshore and offshore) in different operating conditions. This approach outperformed the support vector machine method when dealing with multi-dimensional features sets in terms of avoiding model overfitting.

As seen, ML techniques have been successfully implemented in real case applications. Nevertheless, the ML algorithms adopted in these studies for VSHM frameworks are often referred to as black-box models and they provide little to no information about the decision-making process. The lack of interpretability and understanding regarding how these models detect changes in structural integrity and how DSFs are used by these models to make a certain prediction is one of the reasons hindering the widespread adoption of these techniques for industrial applications. Enabling interpretability in ML-based VSHM frameworks can help building evidence for reliable decision-making in SHM applications. In particular, differentiating between novelties caused due to the influence of EOVs or damage in the structure can help reducing false alarms.

The issue of interpretability when using ML techniques has been addressed previously in fields outside engineering [12]. In particular, the Shapley Additive exPlanations (SHAP) approach proposed by Lundberg et al. [13] has lately attracted attention in the research community. The authors adopted a model agnostic representation of feature importance which is estimated by Shapley values [14, 15] and developed an algorithm to estimate the Shapley values in a computationally efficient way. SHAP is an additive feature attribution method that defines the output of a model as the sum of the real values attributed to each input feature. In other words, the Shapley values represent the marginal contribution of each feature to the prediction of the model. Lundberg et al. [13] have also shown that the Shapley values are the only way to assign feature importance while maintaining two important properties:

Local accuracy: The sum of the feature attributions is equal to the output of the function one is seeking to explain

Consistency: Changes in the model where the impact of a feature increases will never de-

crease the attribution assigned to that feature.

SHAP has been implemented in fields such as medicine and finance. For example, Lundberg et al. [16] used the SHAP technique to understand the outputs of a ML-based system predicting the risk of hypoxaemia during anaesthesia care from electronically recorded intraoperative data. Similarly, in the field of finance, Bussmann et al. [17] employed the SHAP technique to explain the risks associated to peer to peer lending credit in regulated financial services. The interpretable model enables lending entities to classify potential borrowers into risky and non-risky borrowers based on common financial characteristics identified by SHAP as highly influential including earnings before interest and profit before taxes for non-defaulted companies and total assets and shareholders funds for defaulted companies. In both cases, the added interpretability helped understanding the predictions made by the ML models and highlighted the potential of interpretable ML for reasoning in predictive models.

Interpretable ML techniques such as SHAP have also been adopted in engineering applications, albeit less frequently. For instance, Parsa et al. [18] used SHAP to analyse the causes for traffic accidents predicted by a XGboost classification model. By implementing SHAP, the study found that traffic related features such as the vehicle speed had a relative high impact on the prediction of an accident. Furthermore, non-traffic related features such as demographic, network, land use and the weather conditions were identified as important predictors for the occurrence of an accident. This information can facilitate policy decisions related to e.g. speed limits, urban planning, among others.

For SHM applications, Lim et al. [19] estimated different damage levels for several bridges based on monitored conditions of their decks using an XGBoost classification model. Their data was retrieved from the Korean Bridge Management System and corresponded to 142,439 deck inspection records of 2388 bridges. For their analysis, the authors collated 53 variables including identification factors such as region, shape, vehicle weight limit, among others; structural factors such as deck material, thickness and strength, etc.; traffic factors and inspection factors such as age, damage type and condition rating of the damage. Further, SHAP was adopted to capture the influence of the features used for the classification. Their results showed that major influencing

factors influencing damage to bridges include age, average daily truck traffic and vehicle weight limit. Onchis et al. [20] used Local Interpretable Model-agnostic Explanations (LIME) and SHAP in combination to characterize the location and depth of damage in cantilever beams. Through the added interpretability and a novel stability index, the authors provided trust in models prediction. The findings emphasized the benefits of using Shapley values for model interpretability in addition to ML-based techniques and their contribution to the decision-making process for preventive maintenance.

These studies have shown that interpretable ML techniques improved the pattern recognition properties of the predictive models providing supplementary information to select the most adequate predictors (i.e. predictors with the highest contribution). Enabling interpretability in ML-based VSHM frameworks help building evidence for reliable decision-making in SHM applications. However, to the authors' best knowledge, this issue has not been reviewed in the scientific literature, particularly, for interpreting the effect EOVs in continues VSHM for damage detection.

Damage detection in structures can often not be treated as a supervised classification problem, due to the lack of knowledge about the damage. Therefore, this study introduces a novel framework to understand the cause of novelties identified by an MD-based DI in a semi-supervised framework. For this, an XGBoost regression model is trained to learn the relationship between DSFs, EOVs and MD-based DIs. The XGBoost regression is selected due to their performance and low computational time [21]. Furthermore, the process is complemented with the SHAP approach for enhanced understanding of the variabilities in the DIs identified. The proposed framework yields understanding when detecting outliers that correspond to a false alarm or damage in a structure. The framework is first evaluated on a simple lumped mass-springer-damper system with simulated temperature variabilities. The simulation explores and highlights the capabilities of SHAP to interpret the effect of temperature in a dynamic system. Later, the framework is demonstrated on data gathered from an in-operation V27 WT with artificially introduced damage in one of its blades to evaluate and identify the nature of DIs (e.g. EOVs or damage).

The paper is organized such that section 2 introduces the framework for interpretable damage detection followed by a simulated case study in section 3 and an experimental case study on an

operating wind turbine in section 4. Concluding remarks are provided in section 5.

2 METHODOLOGY

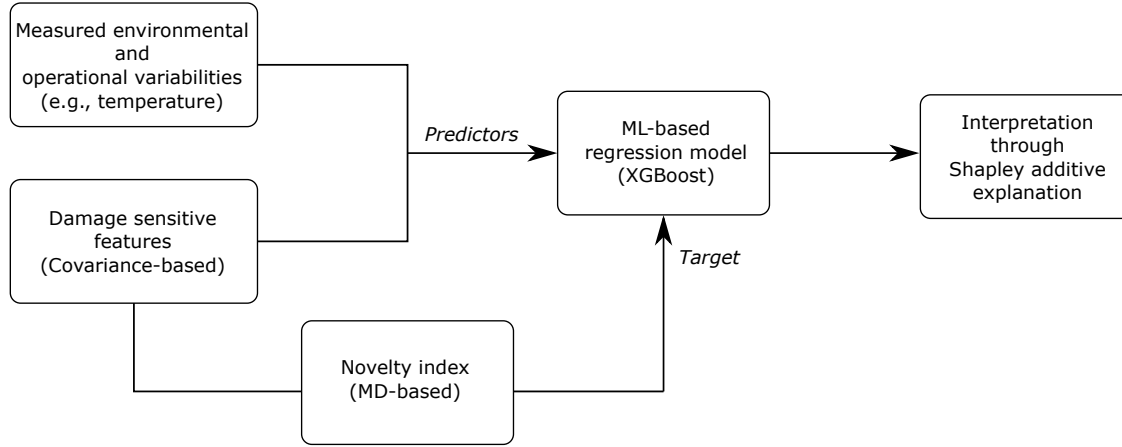


Fig. 1: Proposed flow for interpreting the MD-based DI

This section elaborates on the steps proposed to determine the cause of novelties identified through an MD-based damage detection approach. First, extracted damage sensitive features (DSFs) are used to calculate the MD-based DI. Then, an XGBoost regression model is trained considering the DSFs and measured environmental variabilities as predictors and the MD-based DI as the prediction target. Finally, SHAP is included in the analysis to interpret the XGBoost model and thus interpret the contribution of each DSF and measured environmental parameters to the DI. A schematic representation of the proposed framework is illustrated in Fig. 1.

2.1 Covariance-based damage sensitive features

Using the covariance as a DSF for damage detection has shown promising results in several studies [22–24] and is, therefore, adopted in this study. The covariance of the vibration responses is used to measure the phase and amplitude relationship between the responses at different sensor locations, which characterizes the state of the structure.

2.2 Mahalanobis distance-based damage index

The Mahalanobis distance (MD) measures the distance between a point and a distribution of a multi-dimensional space [25]. Worden et al. [2] introduced the use of the MD as a means of outlier identification for damage detection in structures. To calculate the MD-based DI, DSF vectors $\mathbf{x}_i$ are obtained from the measured responses and gathered into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_M]$, $\mathbf{X} \in \mathbb{R}^{L \times M}$, where L is the length of the DSF vector and M is the number of observations. The MD is calculated as follows:

$$y_i = d_i(\mathbf{x}_i, \mathbf{D}) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu}_D)^T \boldsymbol{\Sigma}_D^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_D)}, \quad (1)$$

where y_i is the MD between the DSF vectors $\mathbf{x}_i$ at observation i and the DSF matrix $\mathbf{D} \in \mathbb{R}^{L \times E}$ with $E \leq M$. $\mathbf{D}$ is the subset of $\mathbf{X}$ used for training. $\boldsymbol{\mu}_D \in \mathbb{R}^L$ is the mean of the observations in $\mathbf{D}$ and $\boldsymbol{\Sigma}_D \in \mathbb{R}^{L \times L}$ is the covariance between observations in $\mathbf{D}$.

2.3 XGBoost decision trees regression model

XGBoost is a scalable machine learning (ML) technique for tree boosting developed by Chen and Guestrin [21]. The algorithm uses nested decision trees to improve efficiency and learning in the model [26]. Whilst decision trees use simple conditional statements to classify the features into target labels based on particular attributes of the features, XGBoost uses multiple decision trees in parallel, thereby optimizing performance and improving the accuracy of the classification. Each decision tree within the XGBoost algorithm learns from the previous decision tree gradually improving the model and building a strong learner as shown in Fig. 2

Let $(\chi_1, y_1), (\chi_2, y_2), \dots, (\chi_i, y_i), \dots, (\chi_M, y_M)$ be a data set with M observations containing the DSFs including measured EOVs $\chi_i \in \mathbb{R}^{\mathcal{L}}$ and the corresponding prediction target $y_i \in \mathbb{R}$. The tree ensemble model is the summation of K additive functions with the estimated prediction $\hat{y}_i$ defined

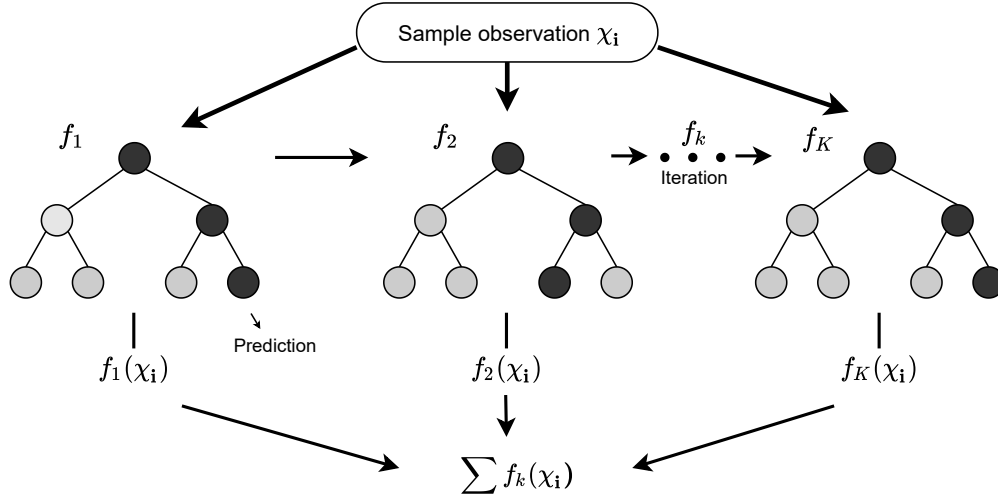


Fig. 2: Basic structure of gradient boosted decision trees

as follows:

$$\hat{y}_i = \sum_{k=1}^K f_k(\chi_i), \quad f_k \in F, \quad (2)$$

where f_k is the independent tree structure of K trees and F the tree space. Different sets of functions are trained by minimizing the following objective function:

$$J^{(t)} = \sum_{i=1}^M l(\hat{y}_i, y_i) + \sum_{k=1}^t \Omega(f_k) \quad (3)$$

where the first regularization term l is a loss function which measures the difference between the prediction $\hat{y}_i$ and the prediction target y_i at iteration step t (or t^{th} tree). The second term Ω is the complexity of the model and is defined as:

$$\Omega(f_k) = \gamma P + \frac{1}{2} \lambda \sum_{j=1}^P w_j^2, \quad (4)$$

where P is the number of leaves in the tree, w_j the score of a leaf, γ the complexity scalar of each leaf and λ the scaling parameter to penalize the complexity. Unlike in general gradient boosting decision trees, for XGBoost the second-order Taylor expansion is taken to the loss function. Using the mean squared error (MSE) as a loss function, the t^{th} tree can be derived as follows:

$$\begin{aligned} J^{(t)} &\approx \sum_{i=1}^M [g_i w_{q(\chi_i)} + \frac{1}{2} (h_i w_{q(\chi_i)}^2)] + \gamma P + \frac{1}{2} \lambda \sum_{j=1}^P w_j^2 \\ &= \sum_{j=1}^P [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma P. \end{aligned} \quad (5)$$

by defining $I_j = \{i | q(\chi_i) = j\}$ as the instance set assigned to the j^{th} leaf. I_j represents all the data samples in leaf node j ; g_i and h_i are the first and second derivatives of the MSE loss function and data points are assigned to the corresponding leaf with the function $q(\cdot)$. Further, by letting $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, optimizing the objective function can be transformed to finding the minimum of a quadratic function. The optimal weight w_j^* for the j^{th} leaf is computed by

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (6)$$

for a given structure $q(x_i)$

$$J^* = -\frac{1}{2} \sum_1^T \frac{G_j}{H_j + \lambda} + \gamma T \quad (7)$$

where Eq. 7 is used as a scoring function to measure for a given tree structure q . To evaluate the best split during the training and determine when the splitting is stopped, the following gain score is calculated:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \quad (8)$$

where the first two terms refer to the summation of the left and right tree respectively and are compared to the score of the original leaf, in third term. The score of the generated leafs is penalized by γ . If the gain is negative, the tree stops growing and returns to the previous split, otherwise it continues improving the model.

In other words, after every split the decision tree (DT) model is evaluated based on the two regularization terms: The first term in Eq. 3 is responsible for prediction accuracy during the training and the second term considers the complexity of the DT. As high tree complexity can cause over-fitting, this is penalized.

2.4 Shapley Additive Explanations for XGBoost decision trees

The Shapley Additive Explanations (SHAP) method is used in this study to interpret the output of the model described in section 2.3. The SHAP method is an additive feature attribution method that defines the output of a model as the sum of the real values attributed to each input feature. Additive feature attribution methods have an explanation model a defined as a linear function of binary features as in the following equation:

$$a(z) = \phi_0 + \sum_{l=1}^L \phi_l z_l \quad (9)$$

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

The SHAP method defines $f_{\mathbf{x}}(S) = E[f(\mathbf{x})|\chi_S]$, where S contains a set of non-zero indexes in z . To compute the feature attribution value ϕ_l for each input feature, SHAP uses the classic Shapley values from game theory as in the following equation:

$$\phi_i = \sum_{S \subseteq N \setminus \{l\}} \frac{|S|!(L - |S| - 1)!}{L!} [f_{\mathbf{x}}(S \cup \{l\}) - f_{\mathbf{x}}(S)], \quad (10)$$

where N is the set of all input features.

To interpret the XGBoost model described in section 2.3, SHAP for tree-based models, i.e.,

TreeShap, as proposed by Lundberg et. al. [27] is implemented.

3 SIMULATED LUMPED SIX DEGREE OF FREEDOM SYSTEM

To evaluate the proposed framework in application to VSHM, let us first consider a simple dynamic system. Figure 3 shows a six degrees of freedom (DOF) lumped mass-spring-damper system consisting of six masses m_1 to m_6 connected to each other by springs with stiffnesses k_1 to k_6 and dampers with the coefficients c_1 to c_1 . The system is fixed to the ground on one of the sides. The mass and damper coefficients are given the same value of $2kg$ and $0.01 \frac{Ns^2}{m}$ respectively.

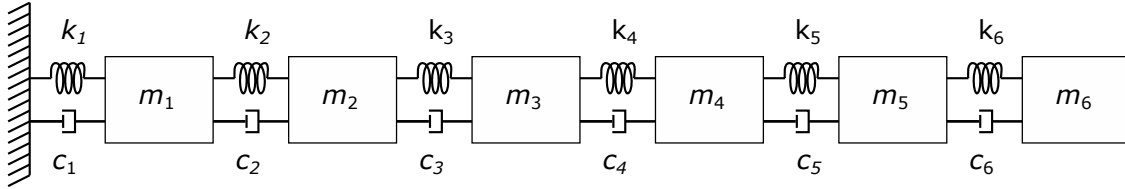


Fig. 3: A six DOF mass-spring system

To model the effect of EOVs on the system's dynamic, we assume the temperature T_i affects the spring stiffnesses. For simplicity, the following rule is used:

$$k_1 = k_2 = k_3 = k_4 = k_5 = k_6 = 0.1 \times T_i, \quad (11)$$

The system is excited by an impulse force applied to first mass and, after the impulse, the masses vibrate freely. The resulting time series with 4350 samples with a sampling frequency of $1321Hz$ are used for calculating the covariance which are used as DSFs (see section 2.1).

In total we simulate 2000 observations. At each observation, the temperature is set according to Figure 4. It ranges from approximately $7.5^\circ C$ to $15.5^\circ C$, where extreme weather events are implemented as sudden temperature drops of $4^\circ C$ between observations 1100-1150 and 1400-1450. To simulate damage, after 1950 observations, the stiffness k_3 is reduced by 5%. For each

observation, the stiffnesses are set according to Eq. 11, and the DSF are calculated. The number of observations per simulated state is summarized in Table 1.

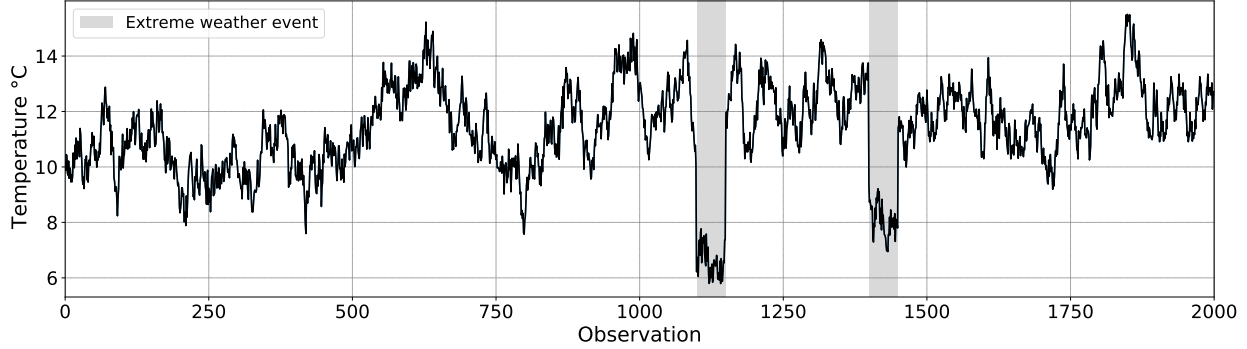


Fig. 4: Simulated temperature field with sudden temperature drops highlighted with a grey background

	Healthy		Damage	Total
	Reference	Testing		
Observations	1250	700	50	2000

Table 1: Partition of data: Simulated data set

3.1 Damage detection: MD-based DI

Following the framework for continues monitoring presented in previous studies [24, 28], the MD-based DI, Eq. 1 is computed with only a subset of healthy observations, in this case the first 1250 are used as a reference for the inverse covariance matrix Σ_D . The DI in Fig. 5 shows an increase for observations 1100-1150 and 1400-1450, where the temperature drops 4°C. Further, the DI of observations 1950-2000 increase due to the 5% stiffness reduction of spring k_3 .

3.2 Interpretability

To understand the difference between novelties caused by the stiffness reduction and the influence of temperature effects, a regression model is built taking the temperature and DSFs

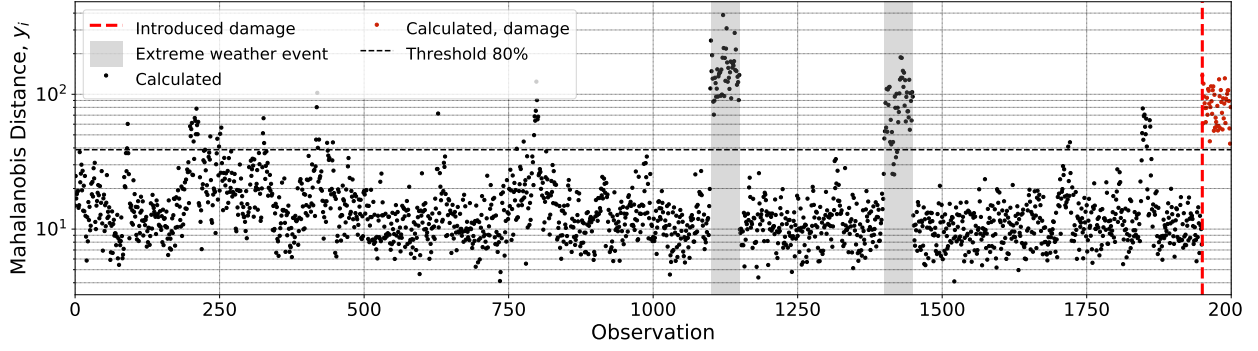


Fig. 5: MD-based damage index

as predictors and the MD-based DI as the prediction target. The XGBoost regression model is trained as described in section 2.3 using observations from the healthy and damaged states of the system. The aim of adopting a regression model in this study is to maintain a high prediction accuracy of the regression model to reproduce the MD-based DI. For this, the hyperparameters (i.e., maximum depth, minimum child weight, gamma, subsample, colsample by tree, alpha for L1 regularization, and learning rate) are identified by a random search. The search is performed until a performance accuracy $R^2 > 0.99$ was reached, where R^2 is square of the sample correlation coefficient between the predicted and the calculated MD. The parameters are summarized in Table 2. For each randomly selected values in a given range a maximum of 1000 iteration is set and the training stops if the mean squared error does not improve for 50 iterations to avoid overfitting.

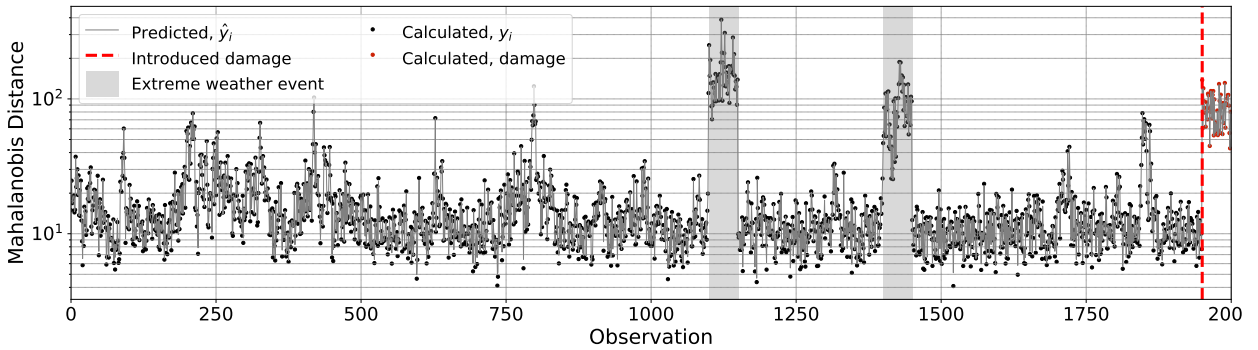


Fig. 6: Predictions from the XGBoost regression model for 6DOF mass-spring-damper system

Hyperparameter	Description	From	To	Interval	6DOF - Best parameters
Maximum depth	The maximum depth of which each tree is built	1	10	1	7
Minimum child weight	The minimum sum of instance weight of all the observations required in a child	1	5	1	3
Gamma	Controls the minimum loss reduction required to make a node split	0	1	0.1	0
Subsample	The number of observations randomly sampled at each tree	0.1	1	0.1	0.4
Colsample by tree	The number of features selected to build each tree	0	1	0.1	0.6
Regularization alpha	L1 regularization value on weights	0.01	0.3	0.01	0.19
Learning rate	Stepsize on updating the weights	0.01	0.3	0.01	0.22

Table 2: Summary of hyperparameters and ranges chosen for the random search

Figure 7a summarizes the Shapley values obtained for the 6DOF lumped mass-spring-damper system XGBoost model for the observations 1 to 1950, while Figure 7b refers to the summary of the observations 1950 to 2000. The features are ranked in descending order. It can be seen that the variance of sensor 3 has the largest marginal contribution on the model's prediction and the covariance of sensor 3 and 5 the lowest. On the horizontal axis, the contribution of the feature can be observed indicating a positive or a negative impact on the models prediction. For example, the temperature feature in Figure 7a shows that low values (colored blue) have a positive impact on the models prediction, i.e. these increase the model's predicted index, whilst lower values have a negative impact, decreasing the model's predicted index. This suggests that low temperatures cause a high MD-based DI and vice-versa. Looking at the last 50 observations in Figure 7b, where the stiffness reduction is introduced, the contribution of the temperature to these predictions drops from ranks 4 to 6. This indicates that the predictions of these observations are not dominated by the influence of the temperature field on a global scale.

Looking at the local Shapley values ϕ_{temp} for the temperature in Figure 8, it can be observed that observations 1100-1150 and 1400-1450 have particularly high Shapley values. This indicates that these observations are caused by temperature. Conversely, observations 1950 to 2000 have a negative impact on the model's prediction indicating that these novelties did not occur due to the influence of temperature.

Figure 9a and 9b show the Shapley values for observations 1100-1150 and 1400-1450 respectively in a global impact overview of the features contribution. In these figures, it is clear that the temperature is the most influential feature for the predictions. The temperature drops of 4°C introduced into the 6-DOF simulated system which were responsible for the increased DI can be identified as false alarms by observing the Shapley values.

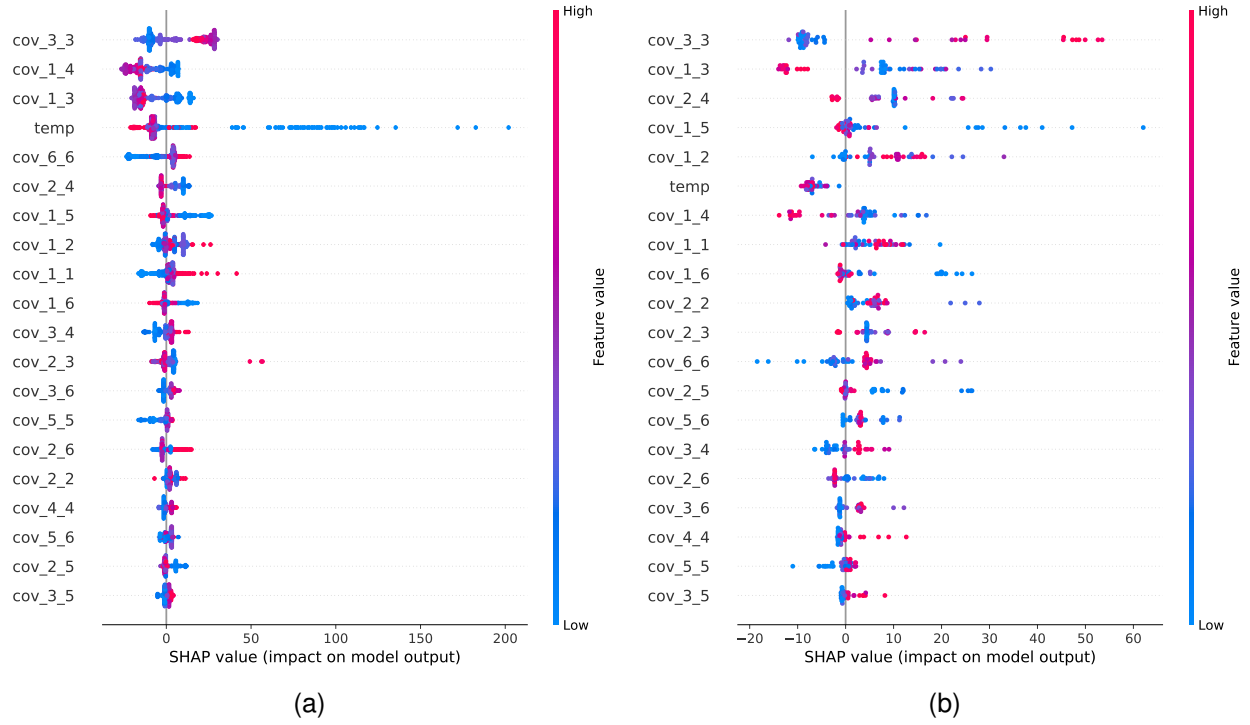


Fig. 7: Summary contribution plot for (a) observations 1 to 1950 with no implemented stiffness reduction and (b) observations 1950 to 2000 with stiffness reduction

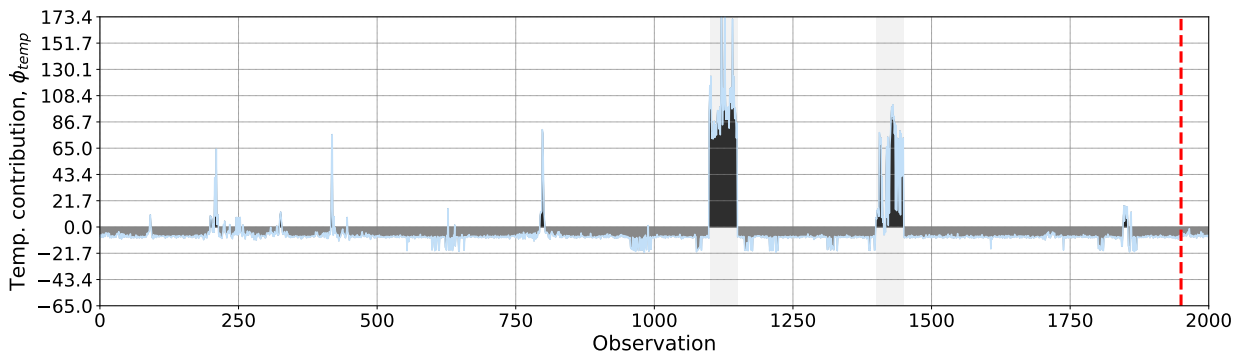


Fig. 8: Shapley explanation for temperature in the 6DOF mass-spring-damper system XGBoost regression model

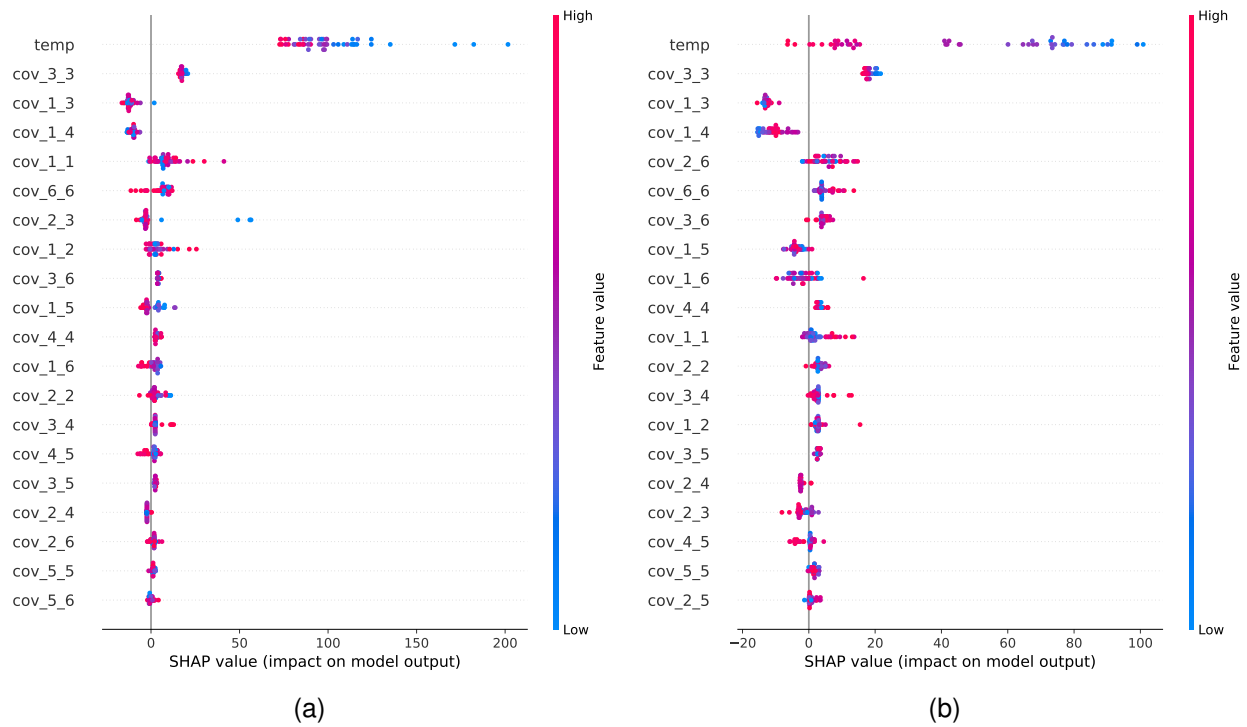


Fig. 9: Summary contribution plot for extreme weather events between (a) observations 1100-1150 and (b) observations 1400 to 1450

4 CASE STUDY: OPERATIONAL WIND TURBINE V27

In this section, we demonstrate the capabilities of the proposed framework to detect damage and explain the DI in a continuous monitoring system installed in an operating wind turbine. The experiment was conducted by Tcherniak et al. [24] between November 28th, 2014 and March 12th, 2015. In the experiment, one of the blades in the operating Vestas V27 wind turbine in Figure 11a is equipped with 11 accelerometers. For this study, the 8 accelerometers shown in Figure 10 placed on the leading and trailing edge are considered. Additionally, the electromechanical actuator shown in Figure 11c was mounted close to the root of the blade to excite the blade. During the monitoring campaign, a total of 24,693 observations were recorded considering a healthy, damaged and a repaired WTB. During the monitoring campaign, artificial damage was introduced by creating a trailing edge opening and gradually expanding it from an initial length of 15cm in damage scenario 1 (see Figure 11b), to 30cm in scenario 2 and 45cm in scenario 3. Finally, the WTB was repaired and further observations were recorded.

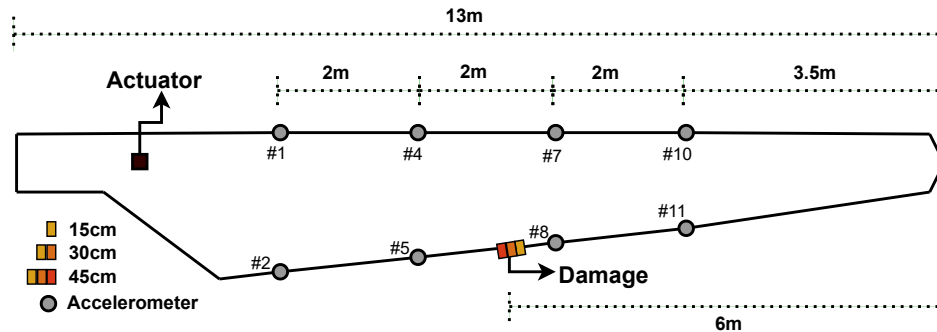


Fig. 10: Sketch of the equipment on the blade of the Vestas V27 wind turbine

The framework is demonstrated on observations corresponding to the 32 rpm operational mode of the turbine. For the purpose of damage detection, the data is classified into two sets, namely healthy and damaged. The former contains solely observations recorded after repairs as there are nearly three times more observations than those recorded prior to the introduction of the artificial damage. Therefore, in this study, the observations considered as healthy are obtained from the repaired observations which consists of 2927 observations. The data classified as

damaged correspond to the observations that were recorded during the three damage scenarios. Further, the observations are filtered by a band-pass filter with cut-off frequencies of 700 and 1200 Hz. A summary of the number of observations in each class is presented in Table 3.

	Healthy		Damage			Total
	Reference	Testing	15cm	30cm	45cm	
Observations	2000	639	66	117	105	2927

Table 3: Partition of observations in the Vestas V27 data set

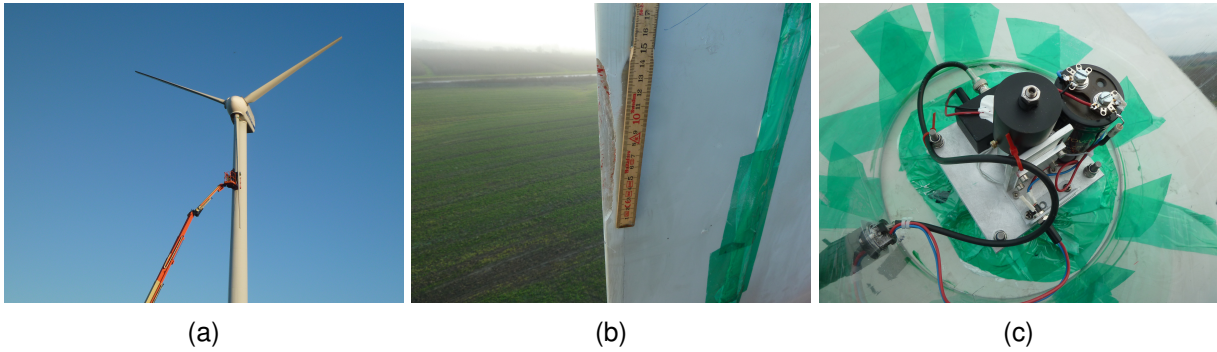


Fig. 11: (a) Vestas V27 wind turbine (b) First damage implemented by a 15cm trailing edge opening around 6m from the tip of the blade (c) Electromagnetic actuator mounted close to the root of the blade for mechanical excitation

4.1 Damage detection: MD-based DI

The DI shown in Figure 12 is computed in the same manner as described in section 2.2 with the obtained covariance-based DSF vectors. The first 2000 observations are used to create a reference to compute the MD as shown in Eq. 1. Outliers in the DI can be identified at around observations 360-675 and 1080-1215 (highlighted in grey). These observations will be the focus of further interpretation. The DI for these observation causes false alarm and, thus, misinterpreted damage.

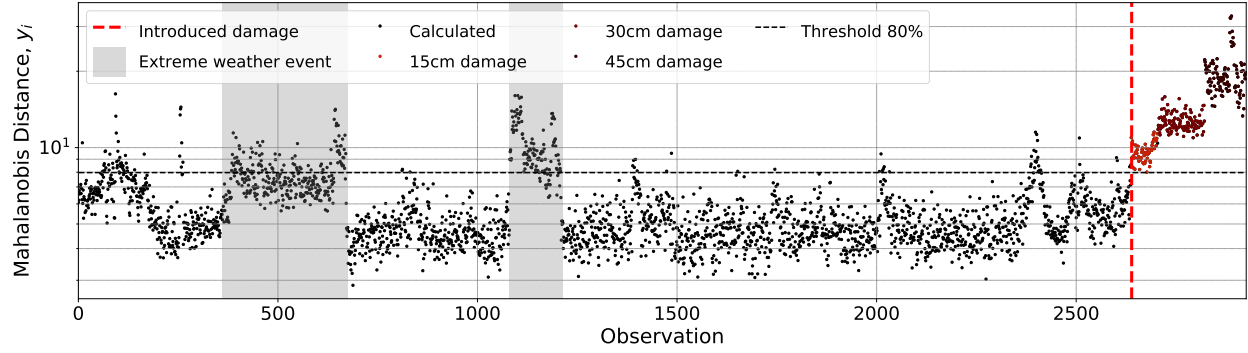


Fig. 12: MD-based DI for wind turbine blade V27

4.2 Interpretability

To understand the potential cause of these novelties and how they differ from the identified damage after observation 2639, the XGBoost regression is built. DSFs and measured EOVs such as temperature, wind direction and wind speed are included. The XGBoost regression model is adopted from section 3.1 with the hyperparameters shown in Table 4. The model's predictions with an $R^2 > 0.99$ are shown in Fig. 13.

Hyperparameter	Description	From	To	Interval	V27 - Best parameters
Maximum depth	The maximum depth of which each tree is built	1	10	1	8
Minimum child weight	The minimum sum of instance weight of all the observations required in a child	1	5	1	1
Gamma	Controls the minimum loss reduction required to make a node split	0	1	0.1	0
Subsample	The number of observations randomly sampled at each tree	0.1	1	0.1	0.6
Colsample by tree	The number of features selected to build each tree	0	1	0.1	0.5
Regularization alpha	L1 regularization value on weights	0.01	0.3	0.01	0.08
Learning rate	Stepsize on updating the weights	0.01	0.3	0.01	0.14

Table 4: Summary of hyperparameters and ranges chosen for the random search

The global influence, i.e., the Shapley values, of the DSFs and EOVs are shown in Figure 14a and 14b. The Shapley values presented for the healthy state of the WTB in Figure 14a show that the variance of accelerometer 7 has the highest influence on the model's predictions, followed by temperature. This indicates that the DI is highly influenced by temperature while the turbine has no artificial damage introduced. We also can identify that temperature values close or below (see Figure 15a) have a positive influence suggesting that low temperatures cause a high MD-based DI. This is consistent with the SHAP results obtained from the simulation in section 3.1, where the

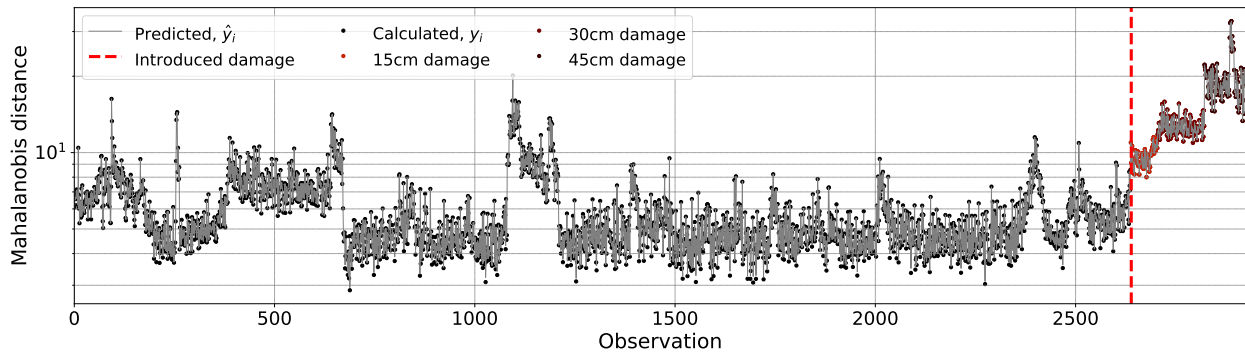


Fig. 13: Predictions from the XGBoost regression model for the V27 MD-based DI

drops in temperature caused an increase in the MD-based DI. EOVs such as wind direction and wind speed have a rather small influence on the model's prediction. Conversely, the significance of the temperature dropped for observations 2639-2927 as shown in Fig. 14b. The majority of Shapley values for observations during the damaged state of the WTB fluctuate around or below 0 indicating a low influence on the model's prediction. This leads to the conclusion that these novelties are not caused by the effect of temperature.

For local interpretability and to gain insight into the contribution of the temperature measurements shown in Figure 15a to the model's predictions, the corresponding Shapley values for each of the observations are shown in Figure 15b. Looking into the previous false alarms identified between observations 360-675 and 1080-1215, it is clear that these outliers are influenced by the positive Shapley values contribution of temperature to the model's outputs. When looking at the observations after 2639, when the artificial damage was introduced, the temperature effect is rather negligible with Shapley values fluctuating around zero.

Focusing on observations 360-675 in Figure 16a and 1080-1215 in Figure 16b, the Shapley values for temperature are dominating throughout all features confirming the influence of EOVs on the model's outputs. Low temperature values have a positive impact on the prediction where higher values have a rather low impact.

Nevertheless, limitations can be found in observations around 2400 where false alarm is present but a rather minor contribution of weather is found slightly reaching a neutral contribution of zero. This indicates, that the model is able to predict these observations but does not use

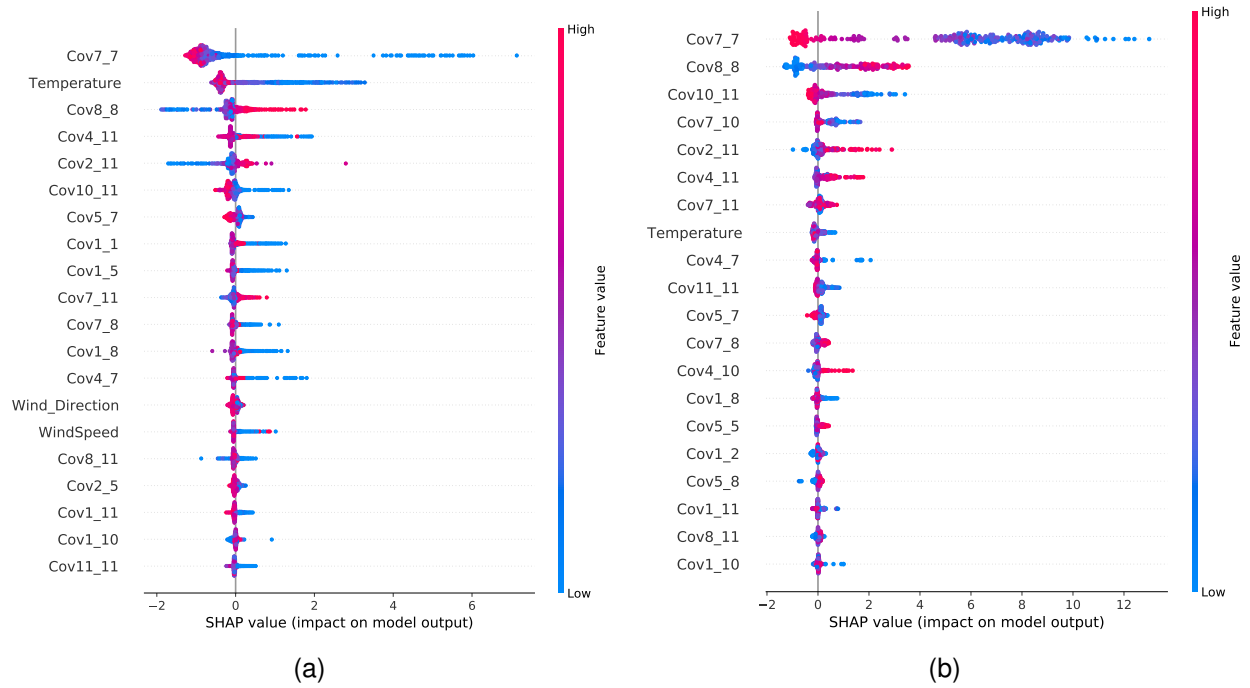


Fig. 14: Summary contribution plot for (a) observations 1-2639 for the healthy state of the structure and (b) observations 2639-2927 with artificially introduced damage

the weather as a significant contributor.

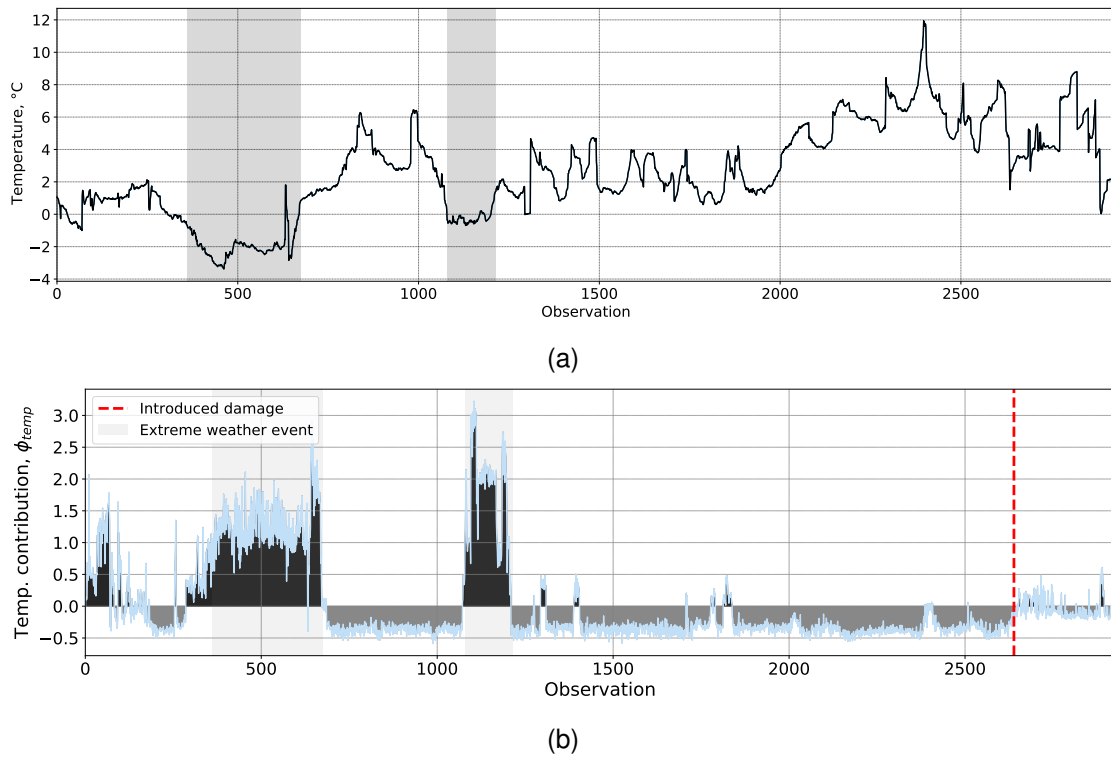


Fig. 15: (a) Measured temperature during the measurement campaign and (b) Shapley explanation for the temperature in the XGBoost regression model for damage detection in the operating V27 WTB

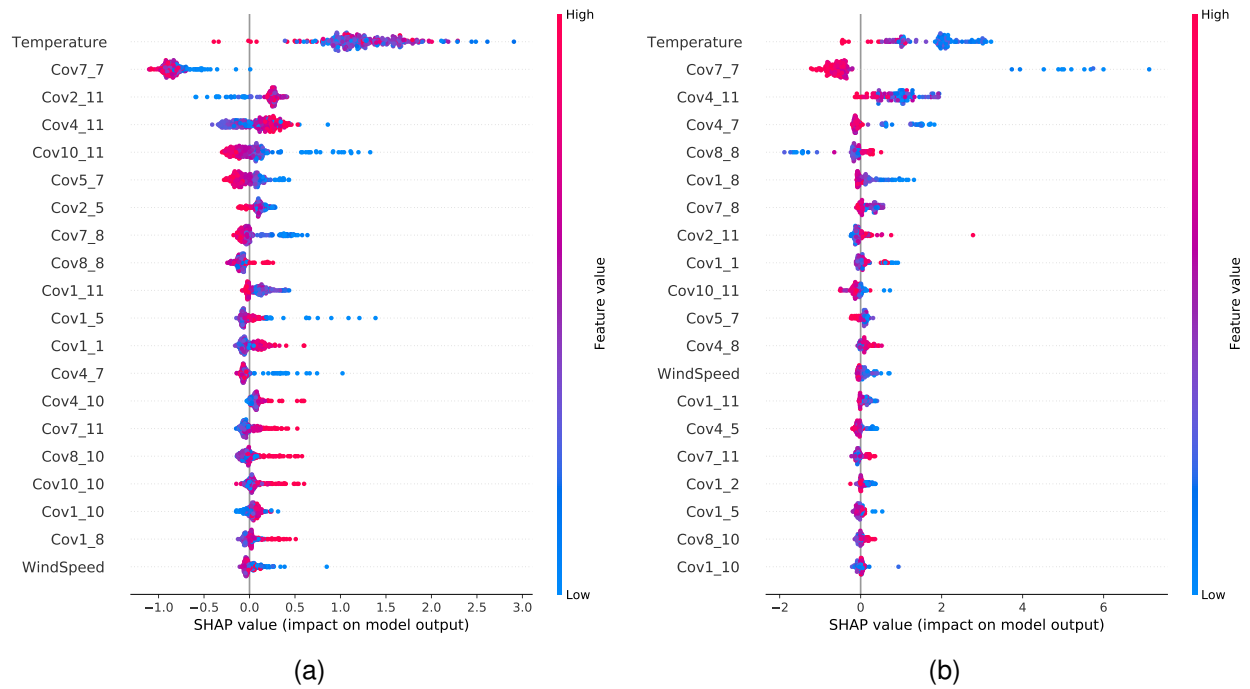


Fig. 16: Summary contribution plot for (a) observations 360-675 where the first significant outliers were identified and (b) observations 1080-1215 where second outliers were identified

5 CONCLUSIONS

This study introduced a framework to distinguish novelties due to the influence of environmental and operational variabilities (EOVs) from those caused by damage in a structure. This framework addresses the lack of interpretability and transparency in ML techniques used for vibration-based structural health monitoring (VSHM). The proposed framework was first evaluated on a 6-degrees-of-freedom lumped mass-springer-damper system with simulated temperature variabilities and, later, demonstrated on a full-scale experiment in an in-operation wind turbine blade with artificially introduced damage.

The results indicate that ML models such as XGBoost can successfully learn the relationship between damage sensitive features (DSFs), EOVs and Mahalanobis distance-based damage indices. The complementary use of interpretable ML, in this case the Shapley Additive exPlanations (SHAP) approach, enabled the determination of the contributions of the DSFs and EOVs to the predictions from the ML model. This process provided transparency and eased the understanding of the influence of EOVs on the damage indices. The SHAP method estimated the marginal contributions of EOVs such as temperature and enabled the identification of false positives. The results suggest that the proposed framework can strengthen the confidence in the decision-making process and contribute to the widespread adoption of ML techniques for industrial applications. Furthermore, the proposed framework can increase reliability and safety and reduce operation and maintenance costs in the wind industry.

REFERENCES

- [1] Farrar, C. R., and Worden, K., 2010. "An introduction to structural health monitoring". *New Trends in Vibration Based Structural Health Monitoring*, pp. 1–17.
- [2] Worden, K., Manson, G., and Fieller, N. R., 2000. "Damage detection using outlier analysis". *Journal of Sound and vibration*, **229**(3), pp. 647–667.
- [3] Sohn, H., Dzwonczyk, M., Straser, E. G., Law, K. H., Meng, T. H.-Y., and Kiremidjian, A. S., 1998. "Adaptive modeling of environmental effects in modal parameters for damage detection in civil structures". In *Smart structures and materials 1998: Smart systems for bridges, struc-*

- tures, and highways, Vol. 3325, International Society for Optics and Photonics, pp. 127–138.
- [4] Kullaa, J., 2011. “Distinguishing between sensor fault, structural damage, and environmental or operational effects in structural health monitoring”. *Mechanical Systems and Signal Processing*, **25**(8), pp. 2976–2989.
- [5] Peeters, B., and De Roeck, G., 2001. “One-year monitoring of the z24-bridge: environmental effects versus damage events”. *Earthquake engineering & structural dynamics*, **30**(2), pp. 149–171.
- [6] Peeters, B., Maeck, J., and De Roeck, G., 2001. “Vibration-based damage detection in civil engineering: excitation sources and temperature effects”. *Smart materials and Structures*, **10**(3), p. 518.
- [7] Figueiredo, E., Park, G., Farrar, C. R., Worden, K., and Figueiras, J., 2011. “Machine learning algorithms for damage detection under operational and environmental variability”. *Structural Health Monitoring*, **10**(6), pp. 559–572.
- [8] 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), p. 2778.
- [9] Santos, A., Figueiredo, E., Silva, M., Sales, C., and Costa, J., 2016. “Machine learning algorithms for damage detection: Kernel-based approaches”. *Journal of Sound and Vibration*, **363**, pp. 584–599.
- [10] 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”. *Journal of Sound and Vibration*, **388**, pp. 154–170.
- [11] 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**, pp. 21020–21031.
- [12] Vilone, G., and Longo, L., 2020. “Explainable artificial intelligence: a systematic review”. *arXiv preprint arXiv:2006.00093*.
- [13] Lundberg, S. M., and Lee, S.-I., 2017. “A unified approach to interpreting model predictions”.

- In Advances in neural information processing systems, pp. 4765–4774.
- [14] Štrumbelj, E., and Kononenko, I., 2014. “Explaining prediction models and individual predictions with feature contributions”. *Knowledge and information systems*, **41**(3), pp. 647–665.
- [15] Roth, A. E., 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.
- [16] 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., et al., 2018. “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery”. *Nature biomedical engineering*, **2**(10), pp. 749–760.
- [17] Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J., 2020. “Explainable ai in fintech risk management”. *Frontiers in Artificial Intelligence*, **3**, p. 26.
- [18] 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”. *Accident Analysis & Prevention*, **136**, p. 105405.
- [19] Lim, S., and Chi, S., 2019. “Xgboost application on bridge management systems for proactive damage estimation”. *Advanced Engineering Informatics*, **41**, p. 100922.
- [20] Onchis, D. M., and Gillich, G.-R., 2021. “Stable and explainable deep learning damage prediction for prismatic cantilever steel beam”. *Computers in Industry*, **125**, p. 103359.
- [21] Chen, T., and Guestrin, C., 2016. “Xgboost: A scalable tree boosting system”. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.
- [22] Movsessian, A., Cava, D. G., and Tcherniak, D., 2021. “An artificial neural network methodology for damage detection: Demonstration on an operating wind turbine blade”. *Mechanical Systems and Signal Processing*, **159**, p. 107766.
- [23] Movsessian, A., Garcia, D., and Tcherniak, D., 2020. “Adaptive feature selection for enhancing blade damage diagnosis on an operational wind turbine”. In Proceedings of the 13th International Conference on Damage Assessment of Structures, Springer, pp. 594–605.
- [24] Tcherniak, D., and Mølgaard, L. L., 2017. “Active vibration-based structural health monitor-

ing system for wind turbine blade: Demonstration on an operating vestas v27 wind turbine”. *Structural Health Monitoring*, **16**(5), pp. 536–550.

- [25] Mahalanobis, P. C., 1936. “On the generalized distance in statistics”. National Institute of Science of India.
- [26] Friedman, J. H., 2001. “Greedy function approximation: a gradient boosting machine”. *Annals of statistics*, pp. 1189–1232.
- [27] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I., 2020. “From local explanations to global understanding with explainable ai for trees”. *Nature machine intelligence*, **2**(1), pp. 2522–5839.
- [28] Movsessian, A., Qadri, B. A., Tcherniak, D., Cava, D. G., and Ulriksen, M. D., 2020. “Mitigation of environmental variabilities in damage detection: A comparative study of two semi-supervised approaches”. In EURO Dyn 2020: XI International Conference on Structural Dynamics.

LIST OF FIGURES

1	Proposed flow for interpreting the MD-based DI	7
2	Basic structure of gradient boosted decision trees	9
3	A six DOF mass-spring system	12
4	Simulated temperature field with sudden temperature drops highlighted with a grey background	13
5	MD-based damage index	14
6	Predictions from the XGBoost regression model for 6DOF mass-spring-damper system	14
7	Summary contribution plot for (a) observations 1 to 1950 with no implemented stiffness reduction and (b) observations 1950 to 2000 with stiffness reduction	16
8	Shapley explanation for temperature in the 6DOF mass-spring-damper system XGBoost regression model	16
9	Summary contribution plot for extreme weather events between (a) observations 1100-1150 and (b) observations 1400 to 1450	17
10	Sketch of the equipment on the blade of the Vestas V27 wind turbine	18
11	(a) Vestas V27 wind turbine (b) First damage implemented by a 15cm trailing edge opening around 6m from the tip of the blade (c) Electromagnetic actuator mounted close to the root of the blade for mechanical excitation	19
12	MD-based DI for wind turbine blade V27	20
13	Predictions from the XGBoost regression model for the V27 MD-based DI	21
14	Summary contribution plot for (a) observations 1-2639 for the healthy state of the structure and (b) observations 2639-2927 with artificially introduced damage	22
15	(a) Measured temperature during the measurement campaign and (b) Shapley explanation for the temperature in the XGBoost regression model for damage detection in the operating V27 WTB	23

16 Summary contribution plot for (a) observations 360-675 where the first significant outliers were identified and (b) observations 1080-1215 where second outliers were identified 24

LIST OF TABLES

1	Partition of data: Simulated data set	13
2	Summary of hyperparameters and ranges chosen for the random search	15
3	Partition of observations in the Vestas V27 data set	19
4	Summary of hyperparameters and ranges chosen for the random search	20