POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Explainable convolutional neural networks for Structural Health Monitoring based on transmissibility functions

TESI MAGISTRALE IN AERONAUTICAL ENGINEERING

Silva, Pedro Henrique, 963883

Advisor:
Francesco Cadini

Co-advisors:
Marc Parziale

Academic year:
2021-22

Abstract: A vast range of fields, such as civil, mechanical and aerospace engineering, have been employing structural health monitoring (SHM) units to enable a safer and more efficient operation of assets. These units rely on algorithms to process the acquired data from the structure of the system to perform damage detection, localization and/or quantification in real time. To do so, within many types of possible data, vibrational signals have been widely and successfully employed, since vibrational properties such as natural frequencies, modal damping and mode shapes of a system depend on its structural properties, which may be subjected to damage-induced changes. More specifically, transmissibility functions (TFs) have created a lot of interest due to the possibility of using them for output-only damage diagnosis algorithms, which simplifies the diagnosis procedure by sparing the need of measuring the input signal. Recently, this kind of structural data used by SHM systems has become more accessible and in very large quantities due to significant technology advancements and cost reduction of sensors, creating a very conducive environment for the application of deep learning models in the SHM field. Although those models have been showing very promising results in terms of prediction accuracy, these usually come at a cost of model interpretability due to their increasing complexity. Indeed, models who lack interpretability are more difficult to be trusted, which is a fundamental characteristic in practical engineering applications such as SHM systems. In order to increase the interpretability of such models, many explainable artificial intelligence (XAI) methods have been proposed, such as the layer-wise relevance propagation (LRP) algorithm. In this work, a convolutional neural network (CNN), a specific type of deep learning model, exploited to process TF data to perform damage detection, localization, and quantification, is interpreted through the use of the LRP algorithm. By considering a numerical case study with different damage scenarios, the relevance values returned by the XAI algorithm were investigated through a statistical analysis. It was observed that the majority of the most

relevant features for the CNN are the most damage sensible and important TFs features, agreeing with the existing physical knowledge.

Keywords: Explainable AI, Structural Health Monitoring, Convolutional neural networks, Transmissibility Functions, Layer-wise Relevance Propagation

1. Introduction

From civil to aeronautical engineering applications, structural components suffer from degradation during their lifecycle. In order to assure the safe operation of a given asset, be it a bridge, a car, or an aircraft, it must be subjected to maintenance during its service life. However, there are different maintenance approaches that can be considered. One of the most common strategies is the preventive or scheduled maintenance, in which the asset is subjected to periodic maintenances. This approach is considered of lower risk because the maintenances are planned to be executed before the end of the designed service life of the maintained components. However, since the components degradation may vary from what is expected from design and the asset is maintained independently from its real condition, components may be repaired or replaced when they potentially still have remaining service life or worse, they may not be repaired in time before they fail, leading to a sub-optimal usage of system parts and a lower availability of it, or even to a catastrophic failure [1]. In order to deal with these issues, a Structural Health Monitoring (SHM) system can be exploited to allow the implementation of a condition-based maintenance strategy [2].

A SHM system is designed by implementing a network of sensors in the monitored structure to continuously acquire data that can then be processed by algorithms to perform damage detection, localization, and/or quantification in real-time. Many SHM methods have been developed, such as the use of electrical resistance-based strain gauges and fiber Bragg grating sensors to indirectly measure damage, but, in particular, vibration-based methods are one of the most widely adopted, since vibrational properties of the structures are known to be damage sensitive. More specifically, these methods rely on the fact that the vibrational properties such as natural frequencies, modal damping and mode shapes of a system depend on its structural properties, which may be subjected to damage-induced changes [3]–[5]. Therefore, by detecting changes in the vibrational characteristics of the structure, algorithms can be used to identify the modifications in the physical parameters and so, perform a structural diagnosis.

Within vibration-based SHM methods, approaches in the frequency domain and time domain have been considered. One of the advantages of the frequency domain methods is that they can be used in more different situations, since the structural dynamic properties, such as mode shapes, modal frequencies and modal damping, are only dependant on the structure itself [4]. With this aim, Frequency Response Functions (FRFs) have shown to be advantageous, due to the fact that they carry a large amount of information over their frequency range [6]–[8]. Despite that, FRFs present a major drawback which is the requirement to not only measure the response of the structure, but also the excitation magnitude, which significantly hampers its computation for applications in which the excitation cannot be easily measured. Hence, one of the possibilities to overcome this is the use of Transmissibility Functions (TFs), defined as the ratio between two spectra outputs (e.g., acceleration or displacement) evaluated at different system degrees of freedom due to an excitation at a given system location. Indeed, in addition to their significant sensitivity to damage, their computation

does not require to perform the measurement of the excitation magnitude, but only the knowledge of the excitation location [9].

In addition to the progresses in the research of damage diagnosis methods, in the recent years, the sensors cost reduction and their technology advancement have made the implementation of complex sensor networks more accessible, enabling the acquisition of a large amount of data that can then be used by SHM systems to continuously monitor the structure condition with an adequate damage diagnosis algorithm. However, in order to make the best use of this big amount of available data, usually a time costly process of signal pre-processing is needed before the data can be used as input by a predictive model. This process requires prior expert knowledge of the data so that the most suited features are selected [10], [11]. Besides the extent work that has been done into this topic, in some cases (e.g., when considering complex and perturbed signals) there is still no unanimity by the research community on what are the best features to be considered to perform damage assessment [11]. Within this context, deep learning has attracted attention due to its capability of tackling this issue by learning complex nonlinear representations, i.e., features, of the raw input data without previous pre-processing [10], [12]. This characteristic has been exploited for several different applications, such as image and speech recognition, object detection, drug discovery and genomics [12]. Following this trend, it has also been applied to SHM systems.

Indeed, with the advancement of deep learning research, mainly fuelled by increasing access to large datasets, powerful computers and development of more efficient learning algorithms, deep learning has achieved fantastic results in different fields, even surpassing human capabilities in some cases. However, such a high performance comes with larger and more complex models. Such complexity ultimately leads to less interpretable models, in which it is difficult to understand the reasoning behind why the model returns an output y for a given input x. Models whose rationale is not easily comprehended by their users are usually harder to be trusted, especially in critical fields such as medicine and autonomous vehicles [13]. Indeed, in addition to sheer accuracy, the ability to enhance human knowledge in the decision-making process and to shed light on correlations present in the data are valuable characteristics of deep learning algorithms applied to practical engineering situations [14]. Inspired by this, the explainable artificial intelligence (XAI) field has seen a significant advance in the recent years. Indeed, it has as objective the development of methods to enhance interpretability of artificial intelligence algorithms, especially deep neural networks. Among the many, one of the promising XAI algorithms is the Layer-wise Relevance Propagation (LRP), which works by propagating the deep neural network prediction backwards and obtaining the relevance (i.e., the importance) of each one of the input features. The LRP algorithm has been successfully applied to several applications, helping to increase the interpretability of models and even identifying model biases [15].

Motivated by this, given that trustworthiness is a key aspect for SHM systems, the present work intends to enhance the interpretability of a deep learning-based algorithm applied to perform damage diagnosis for a SHM system. More specifically, considering a similar case study as proposed in [11], a convolutional neural network (CNN) is considered to detect, locate, and quantify damage of an aluminium structural beam, represented by a numerical model, by using as inputs TFs spectra. Subsequently, the LRP algorithm is applied to better understand the rationale behind the CNN. With this, it was aimed to investigate if the CNN performs the damage characterization based on existing physical intuition and/or if it gives new physical insights into how transmissibility functions can be used to perform the structural diagnosis task.

This thesis is organized as follows: initially, a bibliographic review is presented in Section 2. Subsequently, the methodology, the case study and the results sections are detailed in Section 3. , Section 4. and Section 5. , respectively. Finally, some concluding remarks are reported in Section 6. .

2. Literature review

2.1 Structural Health Monitoring with Transmissibility Functions

Several Structural Health Monitoring (SHM) methods based on different types of sensors, such as electrical resistance-based strain gauges, fiber Bragg grating, or piezoelectric smart layers have been adopted [2], [16]. Within these, vibrational-based methods, which rely on the structure vibration characteristics, to characterize damage [4], have been widely investigated. For example, in [17], a damage detection method using modal frequencies and mode shapes is proposed and tested in a structural beam. In [18], a damage assessment technique by using mode shape sensitivities is proposed and tested with a beam-like structure in laboratory conditions and with data from a highway bridge. In [19], damage identification is performed in a composite beam by using curvature mode shapes and four different methods: absolute difference of curvature mode shape, curvature damage factor, damage index and frequency response function curvature method. Instead of employing specific modal parameters such as modal frequencies, mode shapes and modal damping, other proposed methods have been employing frequency response functions (FRFs), which are preferred due to the fact that they carry information over large frequency ranges [6]–[8]. However, a significant disadvantage of the utilization of FRFs is that they require the measurement of the input excitation, which can be considerably difficult in uncontrolled scenarios.

In view of that fact, output-only methods have sought by the SHM community. Within this context, transmissibility functions (TFs) have shown to be very promising, since they are not function of the input signal magnitude [9]. Moreover, due to their mathematical formulation, TFs are composed by the FRFs zeroes. This aspect has been investigated by [20], being demonstrated that TFs are better indicators of the presence of damage due to the higher damage sensibility of zeroes when compared to poles. By exploiting these positive aspects of TFs, several researches have been accomplished to develop response-only damage diagnosis systems. In [21], a neural network is employed to process TFs magnitude and phase values to perform damage diagnosis. The proposed approach is applied to two different case studies: a physical steel framework and a finite element model of a sandwich beam. The results obtained showed that the neural network was able to perform the structural diagnosis task by identifying the damage induced differences in the TFs. In [22], an auto-associative network is exploited to perform damage detection by using TFs as feature. The main idea was to identify the damage existence through the computation of a novelty index with the output of the auto-associative network, which is trained to reproduce the input data corresponding to the undamaged scenarios seen during its training. The method is demonstrated by using a simulated three degrees of freedom system, in which damage is simulated by reducing the stiffness of one of its elements.

In [23], damage detection is performed by performing outlier analysis with TF data, being considered four case studies. The first case study considers a simulated three degrees-of-freedom system, with damaged conditions being simulated by introducing a stiffness reduction, the second case considered experimental and simulated vibration data from a gearbox with local damage in a spur gear, the third considered experimental and simulated lamb-wave data from two damaged composite plates and the last one considered experimental data from a ball-bearing with different

fault conditions. One important aspect noted is that the features related to the peaks of the TFs in the considered frequency range are the ones that render the greatest contribution to the discordancy of the outlier. In [24] a damage detection procedure is developed by identifying the difference between healthy and unhealthy conditions through the use of an integral damage indicator based on the difference between healthy and damaged TFs.

In [25] it is performed the experimental validation of a SHM system whose damage detection algorithm exploits TFs. In fact, three approaches are considered for this task: outlier analysis, density estimation and an auto-associative neural network. The experimental validation is performed in plate with stiffeners, simulating an aircraft wingbox. Subsequently, in [26], [27], the authors experimentally validated the outlier analysis approach presented in [25] in a real aircraft wing. In [28], TFs computed from data obtained from wired piezoelectric arrays of accelerometers are used to perform damage detection and localization through an approach named structural diagnostics using non-linear analysis (sDNA). In [29], TFs are used to detect and locate damage on a section of a wind turbine blade made of fiberglass through the use of an integral damage indicator that considers the difference between the healthy and damaged TFs. In [30], a damage indicator based on the correlation between healthy and damaged TFs is used to perform damage detection and quantification and it is compared to a similar damage indicator based on FRFs, which is shown to be significantly less sensitive than the TF-based one. In [31], it is investigated how operational, sensing, environmental and computational variabilities influence a TF-based damage indicator. Then, improvements, such as limiting the frequency ranges or compensation routines, are proposed to make the SHM system more robust with respect to these variabilities. In [32], it is proposed to feed TF data to an artificial neural network to detect, locate and quantify damage in a bridge structure.

2.2 Deep Learning and Explainable Artificial Intelligence

As mentioned in previous paragraphs, a wide range of algorithms have been considered to perform damage characterization. Within this context, deep learning models have recently been standing out, driven by an increase of available data and computational power. In fact, due to an advancement of sensors' technology and their cost reduction, structural sensing has become more accessible and with a greater amount of available data, making deep learning models strong candidates for SHM applications.

Deep learning models are basically composed of an input layer, hidden layers, and an output layer. As the model receives data through the input layer, it is consecutively transformed as it passes through the hidden layers. Each hidden layer uses as input data the output from the previous layer. By exploiting consecutive non-linear transformations, deep learning models are able to learn complex functions and extract high-dimensional features, which is the reason why this kind of model thrives in conditions in which very large datasets are available. Since the extraction of these features is learned by the model during its learning stage, there is no need to perform feature engineering, which requires domain expertise, being one of main advantages of deep learning models over conventional machine-learning techniques [12].

Exploiting these aspects, deep learning models have been applied in several fields, such as speech recognition [33]–[36], image recognition [37], [38] and autonomous driving [39]. Indeed, they have also been exploited in the SHM field, as highlighted in the state-of-the-art reviews presented in [40] and [41].

In addition to that, some researches have exploited a specific type of deep neural network architecture called convolutional neural networks (CNN). When compared to common fully connected layers, CNNs have the capability of learning local feature representations, require a smaller quantity of learnable parameters and converge faster [12], [42]–[44].

Taking advantage of these aspects, in [45] a framework called DeepSHM is presented, which considers a CNN to process ultrasonic guided waves data to perform damage identification, localization and quantification. Also, in [46], a CNN is used to identify and locate damage by processing acceleration time responses of a beam structure. In order to train the CNN, a finite element model, tuned according to experimental data, is used to generate the training data. In addition to that, in [47] a CNN is also used to perform damage identification and localization through the analysis of raw vibration measurements. In this case, model order reduction techniques are applied to generate the training data through physics-based models with lower computational cost.

Exploiting the TFs characteristics detailed in the previous subsection, in [11] a deep CNN-based approach for damage detection, localization and quantification is proposed. Through the employment of a CNN, raw TFs data is processed, thus not requiring a previous feature engineering step to extract spectral lines or antiresonant frequencies as in other previous works. More specifically, it is used as input data grey scale images which translate the TFs logarithmic magnitude across the considered frequency range, in this case 0 Hz – 2000 Hz. The approach is applied to two different case studies: a mass-spring system with eight degrees of freedom and a structural beam. In order to simulate damage, the structures' elements have their stiffnesses reduced. For both case studies the proposed approach rendered satisfactory results.

In fact, one of the reasons that made possible to obtain remarkable results in several fields was the increase of the deep learning models complexity. However, this increase of complexity came at a cost of models with reduced capabilities of returning interpretable outputs. Indeed, machine learning models can be classified in two groups. First, there are white-box models or glass-box models, such as linear models and decision trees, which are able to produce results that are more easily interpreted but are not able to attain prime performance. Then, there are the black-box models, such as deep learning models, that present state-of-the-art performance at the expense of prediction interpretability [48]. Due to the lack of clarity in its decision-making process, black-box models are more difficult to be trusted, which is a critical aspect for fields such as healthcare, autonomous vehicles, and SHM. In this sense, interpretable models are extremely valuable, since they make easier to comprehend its decision-making reasoning and cause-and-effect relationships between outputs and inputs, increasing the model trustworthiness and also presenting the ability to enhance human expertise in the decision-making process and revealing the correlations it perceives in the data, which are important features sought when applying artificial intelligence to practical engineering of systems [14].

In view of that, a great effort has been made in the field of explainable artificial intelligence (XAI), which aims to enhance the comprehension of the rationale of artificial intelligence models. An example is DARPA's XAI program, which is motivated by the critical need of explainable models for the United States Department of Defense [49]. In order to achieve explainable AI systems, many methods have been considered, each of them with their specific characteristics. For instance, it is possible to implement intrinsically interpretable models (i.e., white-box models) or to choose to explain a black-box model after it has been already built, defined as post-hoc explanation. Moreover,

it can be chosen between local methods, which explain a single prediction at time, or global methods, which explain the whole model. In addition to that, interpretability methods can be divided between model specific, when they can only be applied to a single model or group of models, and model agnostic, when they can be applied to any model. Finally, the interpretability methods also depend on the data type considered (e.g. image, tabular, text,) [48], [50].

A well-known post-hoc model agnostic XAI method is the Shapley Additive explanations (SHAP), which is inspired in game-theory. It proposes SHAP values, which evaluates the features' importance based on how each feature contributes to the model achieving the analyzed output from the expected model prediction [51]. For instance, SHAP has been applied to explain AI models used in different applications, such as accident detection [52], metallurgical processes [53], power systems control [54], and credit risk assessment [55], [56]. LIME, local interpretable model-agnostic explanations, is another popular model agnostic XAI method. LIME approach consists in creating a glass-box model, such as a decision tree, that is able to locally approximate the analyzed black-box model for the interpreted prediction [57]. LIME has been applied to interpret solar photovoltaic power generation forecasting models [58], healthcare [59], and credit risk assessment [60]. Within the post-hoc model specific algorithms, it is possible to divide between methods that follow an approach based on sensitivity analysis and those that opt for a decomposition approach. Sensitivity-based approaches seek to explain the prediction based on the effect of infinitesimal perturbances in the input, leading to an explanation of a local variation of the function that represents the model. On the other hand, decomposition approaches explain a given prediction by redistributing the value of the prediction of the model back to the input variables, in such a way that the prediction function value is conserved, i.e., the sum of the redistributed terms are equal to the prediction value. By doing this, the decomposition approach is able to explain the whole function that represents the model [61]. Among the methods that follow a decomposition approach, the Layer-wise Relevance Propagation (LRP) is one that has shown to be promising. By considering specific local propagation rules, the LRP, which can be applied to deep neural networks with different inputs such as text, images, and videos, works by propagating the model prediction backwards in the neural network in order to obtain the relevance of each input feature for the explained prediction [15]. For instance, the result of the application of the LRP algorithm to explain the prediction made by a CNN used to classify images as "dogs" or "cats" is a vector with the relevance score of each pixel of the input image, which indicates how much each pixel contributed to the prediction.

The LRP algorithm has been successfully employed in several applications. In [62], LRP is used to explain 1D-CNNs in two difference case studies: credit card fraud detection and telecom customer churn prediction. In [63], a Long-Short Term Memory (LSTM) model, a specific architecture of recurrent neural network, is used to perform therapy prediction. Subsequently, LRP is used to explain the model in order to obtain the most relevant features. It is shown that the highlighted features by the LRP are agree to clinical knowledge and guidelines. Moreover, LRP has been considered to explain deep learning models used to perform diseases diagnosis. For instance, in [64], LRP is used to explain predictions performed by a CNN to process structural magnetic resonance imaging data to detect Alzheimer's disease. It is shown that the relevance heatmaps produced by the LRP algorithm correlate well with scientific knowledge. In addition to that, in [65] LRP is exploited to explain the predictions of a CNN used for diagnosing multiple sclerosis. Moreover, LRP has been used in [66] to distinguish schizophrenia patients from healthy individuals. LRP has also been exploited to explain a deep learning model used to investigate intermolecular noncovalent interactions [67]. Another example of LRP application appears in [68], where it was used to explain

a deep learning model used for speech recognition. Within the SHM field, instead, the LRP has been employed in [69] to interpret the predictions performed by a CNN exploited to process time-frequency spectra images of vibration signals to perform fault diagnosis on an induction motor. In addition to that, in [70] LRP is used to interpret a CNN used in a vibration-based damage detection system for machines with variable rotation speed. LRP is also applied in [71] to interpret a deep neural network used to detect and quantify damage in buildings subject to earthquakes. Moreover, in [72] LRP is exploited to interpret CNNs used to detect and locate damages by processing ultrasonic guided waves.

3. Methodology

3.1 Transmissibility Functions

Given a stable dynamic system subject to a single input, the relationship between the response x_i^k at degree of freedom i due to an excitation f_k applied in a generical degree of freedom k is given by Equation (1)**Error! Reference source not found.**:

$$x_i^k(\omega) = H_{ik}(\omega) f_k(\omega) \tag{1}$$

where $H_{ik}(\omega)$ is the FRF between degrees of freedom i and k.

A TF $T_{i,j}^k$ which relates the responses of degrees of freedom i and j due to an excitation at degree of freedom k is defined as shown in Equation (2)

$$T_{i,j}^k(\omega) = \frac{x_i^k(\omega)}{x_j^k(\omega)} \tag{2}$$

in which x_i^k and x_j^k are the responses of degrees of freedom i and j due to an excitation at degree of freedom k, respectively. It should be noticed that the outputs here considered can be either the displacements, velocities, or accelerations.

By substituting Equation (1) into Equation (2), the TF can be rewritten as shown in Equation (3).**Error! Reference source not found.**

$$T_{i,j}^k(\omega) = \frac{H_{ik}(\omega) f_k(\omega)}{H_{jk}(\omega) f_k(\omega)} = \frac{H_{ik}(\omega)}{H_{jk}(\omega)} \tag{3}$$

Which highlights the fact that the TFs are not function of the input signal magnitude, but only on the excitation location.

A generic FRF between two points x and y can be written as show in Equation (4).

$$H_{xy}(\omega) = \frac{n_{xy}(\omega)}{d(\omega)} \tag{4}$$

Where $n_{xy}(\omega)$ characterizes the FRF zeroes and $d(\omega)$ characterizes the system poles. By substituting Equation (4) into Equation (3), Equation (5) is obtained, which shows that TFs are the ratio of the

FRFs numerators, meaning they are only function of the FRFs zeroes, giving them a higher damage sensibility, as investigated in [20].

$$T_{i,j}^{k}(\omega) = \frac{n_{ik}(\omega)}{n_{jk}(\omega)} \tag{5}$$

In the considered case study, in which a structural beam is analyzed, TFs considering different nodes are collected in a matrix form which is then processed by a CNN without any previous feature extraction to perform the damage diagnosis.

3.2 Convolutional Neural Networks

A CNN, whose scheme can be seen in Figure 1, is normally composed of three main types of hidden layers (i.e., the layers between the input and the output): convolutional layers, pooling layers and fully-connected layers.

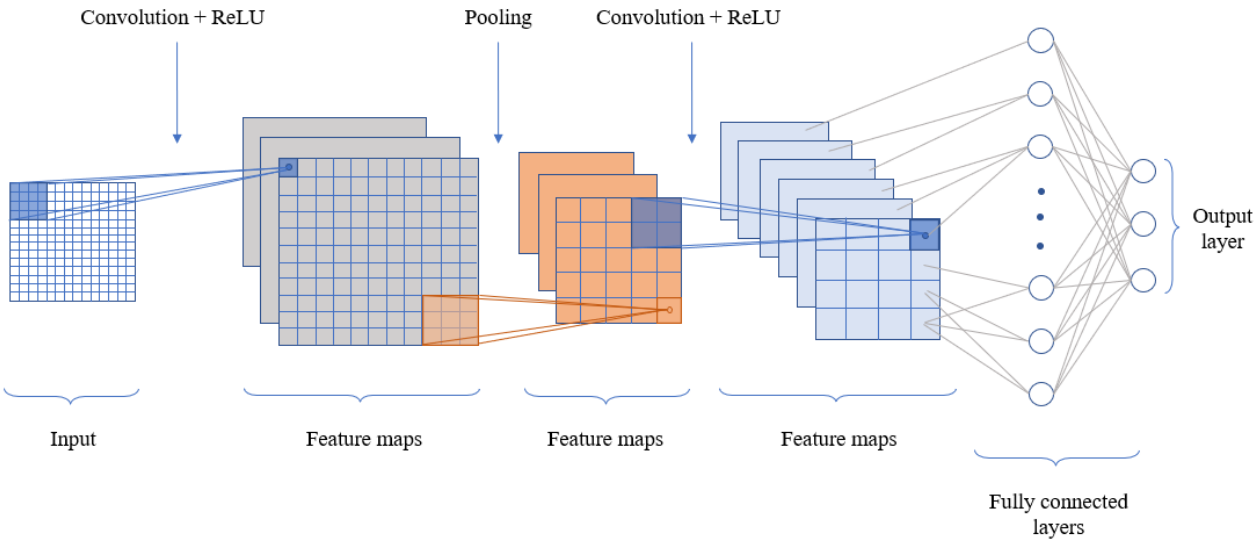

Figure 1: Typical CNN scheme.

As defined in [44], the convolution operation can be mathematically described as in Equation (6):

$$z_{i,j,k}^{l} = {w_k^l}^T x_{i,j}^{l-1} + b_k^l \tag{6}$$

where $z_{i,j,k}^{l}$ is the feature value at location (i,j), $x_{i,j}^{l}$ the input patch also centered at (i,j), and w_k^l and b_k^l are the convolution kernel and bias term of the k-th feature map, respectively, all of them for the considered generical l-th layer. After performing the convolution, the $z_{i,j,k}^{l}$ term passes through an activation function, resulting in Equation (7):

$$a_{i,j,k}^{l} = a(z_{i,j,k}^{l}) \tag{7}$$

For instance, considering a ReLU, it is obtained Equation (8):

$$a_{i,j,k}^{l} = \max(0, z_{i,j,k}^{l}) \tag{8}$$

The pooling operation is defined as in Equation (9):

$$y_{i,j,k}^{l} = \text{pool}\left(a_{i,j,k}^{l}\right), \forall (i,j) \in R_{ij} \tag{9}$$

where R_{ij} is a local region around location (i, j) where the pooling operation is performed and pool() denotes the pooling operation. Two main types of pooling operations can be considered: max pooling and mean pooling.

Finally, the fully connected layers are defined mathematically as in Equations (10) and (11).

$$z_j^l = \sum_i a_i^{l-1} w_{ij}^l + b_j^l \tag{10}$$

$$a_j^l = a(z_j^l) \tag{11}$$

where a_i^{l-1} is the activation of neuron i of the $l-1$-th layer, w_{ij}^l is the weight of this activation to the neuron j of the l-th layer, b_j^l is the bias term for neuron j of the l-th layer, z_j^l is the linear combination of the activations of the neurons of the l-th layer and a_j^l is the activation of neuron j of the l-th layer.

The weights and biases are determined through the use of gradient descent and the backpropagation algorithm. More specifically, in order to reduce the computation cost, the Stochastic Gradient Descent (SGD) algorithm is usually considered instead of the gradient descent. The SGD algorithm works by computing the gradient only considering a mini-batch of the training data, instead of considering the whole dataset [44]. Moreover, many algorithms have been proposed to improve SGD, such as Stochastic Gradient Descent with momentum (SGDm) Root Mean Square Propagation (RMSProp), Adaptative Gradient (AdaGrad) and many others [73]. Finally, with respect to the loss function to be optimized, several options are also available. For instance, for regression tasks, a common loss function is the mean squared error (MSE).

For the considered case study, the TFs are collected into images that are used as input for a CNN, which process it and then returns an output vector which states the damage percentage of each of the element of the structural beam.

3.3 Layer-wise Relevance Propagation

The LRP algorithm works by propagating the neural network prediction backwards through its layers. In order to do so, a propagation rule defined as in Equation (12) is applied.

$$R_x = \sum_y \frac{z_{xy}}{\sum_x z_{xy}} R_y \tag{12}$$

where R_x is the relevance score of neuron x in the layer l, R_y is the relevance score of neuron y in the layer $l+1$ and z_{xy} quantifies how much neuron x contributes to make neuron y relevant. For

instance, for a deep neural network with fully connected layers as defined in the previous Section (i.e., Equations (10) and (11)), the LRP propagation rule becomes

$$R_x = \sum_y \frac{a_x}{\sum_x a_x} R_y \tag{13}$$

The presence of the denominator factor enforces a conservation property in the layers to the propagation rule given by Equation (14).

$$\sum_x R_x = \sum_i R_i \tag{14}$$

This conservation property can be extended to a global conservation property, as shown in Equation (15)

$$\sum_i R_i = f(x) \tag{15}$$

which states that the sum of the relevance of the nodes of the input layer (i.e., $\sum_i R_i$) is equal to the output of the neural network (i.e., $f(x)$). A scheme of the LRP procedure can be seen in Figure 2.

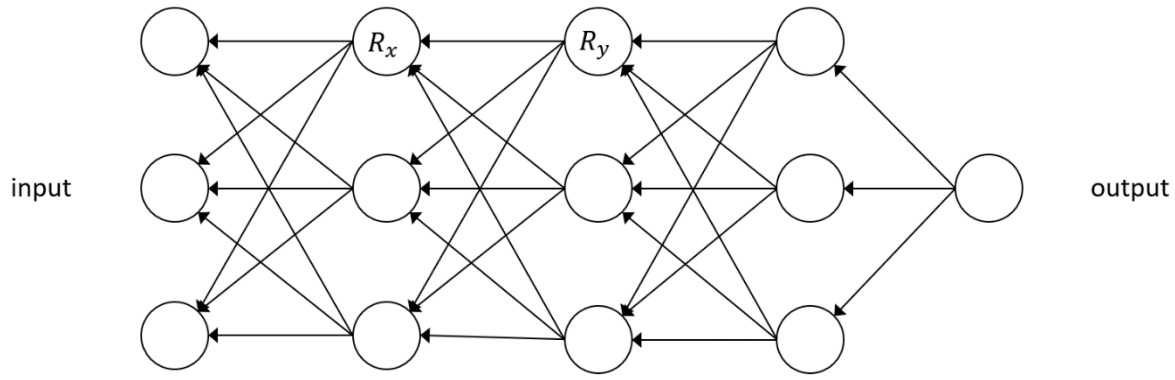

Figure 2 – LRP Scheme highlighting how the relevance is propagated from the output to preceding neurons in a generic neural network.

Many propagation rules have been proposed as improvement of the rule described by Equation (12). One of these is the Epsilon propagation rule, which is given by Equation (16).

$$R_x = \sum_y \frac{z_{xy}}{\epsilon + \sum_x z_{xy}} R_y \tag{16}$$

Through the addition of a small positive term ϵ in the denominator, this propagation rule is able to filter the contributions from neuron y that are weak, leading to explanations that are less noisy and sparser with respect to the input features [15].

In the proposed case study, the LRP is used to interpret the output of the CNN used to perform the damage localization and quantification, highlighting the most relevance features from the input.

4. Case study

A case study similar to the one presented in [11], [74] is considered, in which a CNN is exploited to diagnose damage (i.e., damage detection, localization, and quantification) of a structural beam by processing TFs. Then, the LRP algorithm is employed to interpret how the CNN is making its predictions.

The structural beam considered is an aluminum beam with a length of 1 m and a rectangular cross section of width 25 mm and height 10 mm. In particular, it is modelled with 20 Timoshenko beam elements with equal length of 50 mm.

Under free-free conditions, the beam response is simulated as subjected to a shaker exciting its left tip. The computation is performed with a frequency resolution of 1 Hz in the frequency range from 0 Hz to 2000 Hz. Then, in order to compute the TFs, the response of the translational degrees of freedom of the flexural vibration response is considered. More specifically, 10 TFs are computed. Each TF is computed by taking into consideration the response of node 1, which is the node that is excited by the shaker and chosen as reference one, and an additional node. For instance, TF 1 is computed by considering the responses of node 1 and 3, i.e. $T_{3,1}^{1}$, TF 2 the responses of node 1 and 5, i.e. $T_{5,1}^{1}$ and so on, with TF 10 considering the responses of node 1 and 21, i.e., $T_{21,1}^{1}$. For a more concise notation, the TFs will be called by their respective index as explained. Both damaged and undamaged scenarios are simulated in order to create the dataset used to train the CNN. The approach used to simulate damage in one element is the reduction of a percentage ($d\%$) of its stiffness. In particular, all damage scenarios considered that only one element of the beam is damaged.

In order to process the 10 TFs data with the CNN and since they are computed with frequency resolution of 1 Hz between 0 Hz and 2000 Hz, the logarithmic magnitudes of the 10 TFs are grouped in a matrix T with dimensions 10x2000, in which the first line of the matrix contains the logarithmic magnitude of the TF 1, the second line the logarithmic magnitude of TF 2, and so on. Subsequently, this matrix T can be represented as a gray-scale image, which can then be processed by the CNN. In Figure 3 all 10 TFs for an undamaged scenario are plotted and in Figure 4 a gray scale image originated by the same TFs is shown. Note that this gray scale image is originated by an expanded matrix T_{exp} of dimension 2000x2000, in which logarithmic magnitude of TF 1 are contained from lines 1 to 200, the logarithmic magnitude of TF 2 from lines 201 to 400, and so on. Such approach is considered just to enhance the visualization of the image that is being used as input by the CNN.

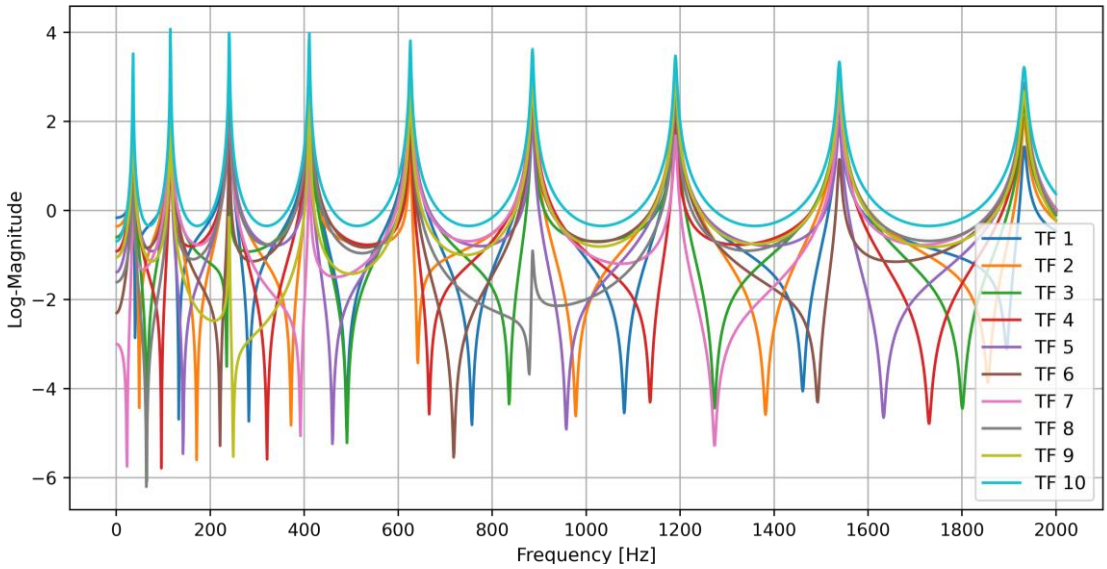

Figure 3 – Healthy TFs plot.

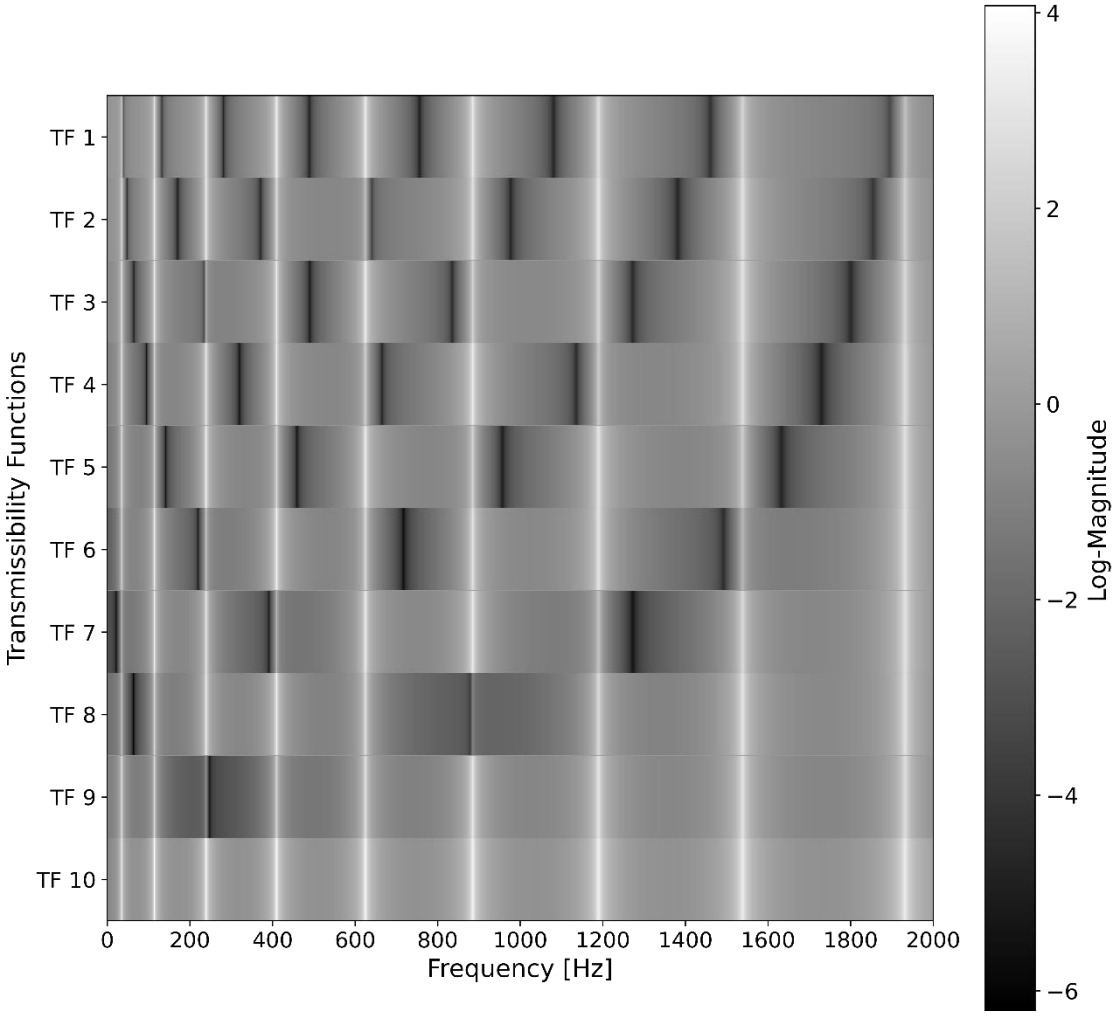

Figure 4 – Gray scale image representing the healthy TFs.

The described gray scale image with the collected TFs is then the input of a CNN, whose architecture's scheme is seen in Figure 5, used to provide the structural health state of the beam. In particular, its architecture consists of two convolutional layers and two fully connected layers. More specifically, the first convolutional layer has a kernel of dimensions (10,1) with 32 filters, mainly responsible to process features across all TFs. The second convolutional layer has a kernel of dimensions (1,5) with 64 filters and its role is that of processing the features across the frequencies. Both convolutional layers, as well as the first fully connected layer, whose size is (1024,1), use ReLU as activation function. Finally, the last fully connected layer, which is the output layer, does not have any activation function and it returns a vector of size (20,1) with the predicted damage percentage $d_n^{CNN}\%$ of each n-th beam element, in which element $n = 1$ is the element composed by nodes 1 and 2, element $n = 2$ is the element composed by nodes 2 and 3, and so on. The MSE is used as loss function.

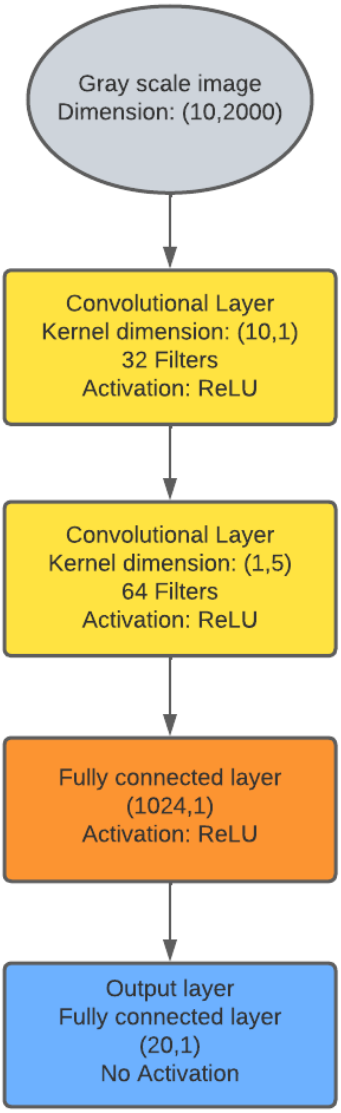

Figure 5 – CNN architecture scheme.

To train the CNN, a dataset of 15000 gray scale image is considered. In particular, 30% of the dataset consists of undamaged scenarios, whereas the rest consists of damaged scenarios. The latter are obtained by changing the location of the damaged element p, which is sampled from the uniform probability mass distribution $U(1,20)$ with unitary step, and the damage extent $d\%$ introduced, that is sampled from the uniform probability mass distribution $U(5,40)$ (%), in this case with a 5% step. In addition to that, a numerical noise is introduced in the computed responses in order to add variability to the dataset and also simulate sensors noise. At each frequency, the noise percentage δ is sampled from the uniform probability mass distribution $U(0,30)$ (%) with a 10% step. Subsequently, the CNN performance is assessed through a validation dataset of 3000 gray scale images, of which approximately 20% are undamaged scenarios. In a similar way of the training dataset, the location of the damaged element p is sampled from the uniform probability mass distribution $U(1,20)$ with unitary step, but the damage extent $d\%$ introduced is sampled from the uniform probability mass distribution $U(7.5, 37.5)$ (%), with a 5% step. Moreover, the noise is applied to the data in a similar way as done for the training dataset. In Figure 6, a gray scale image representing the TFs of a scenario in which element 12 presents 12.5% of damage (noise with $\delta = 30\%$) is shown.

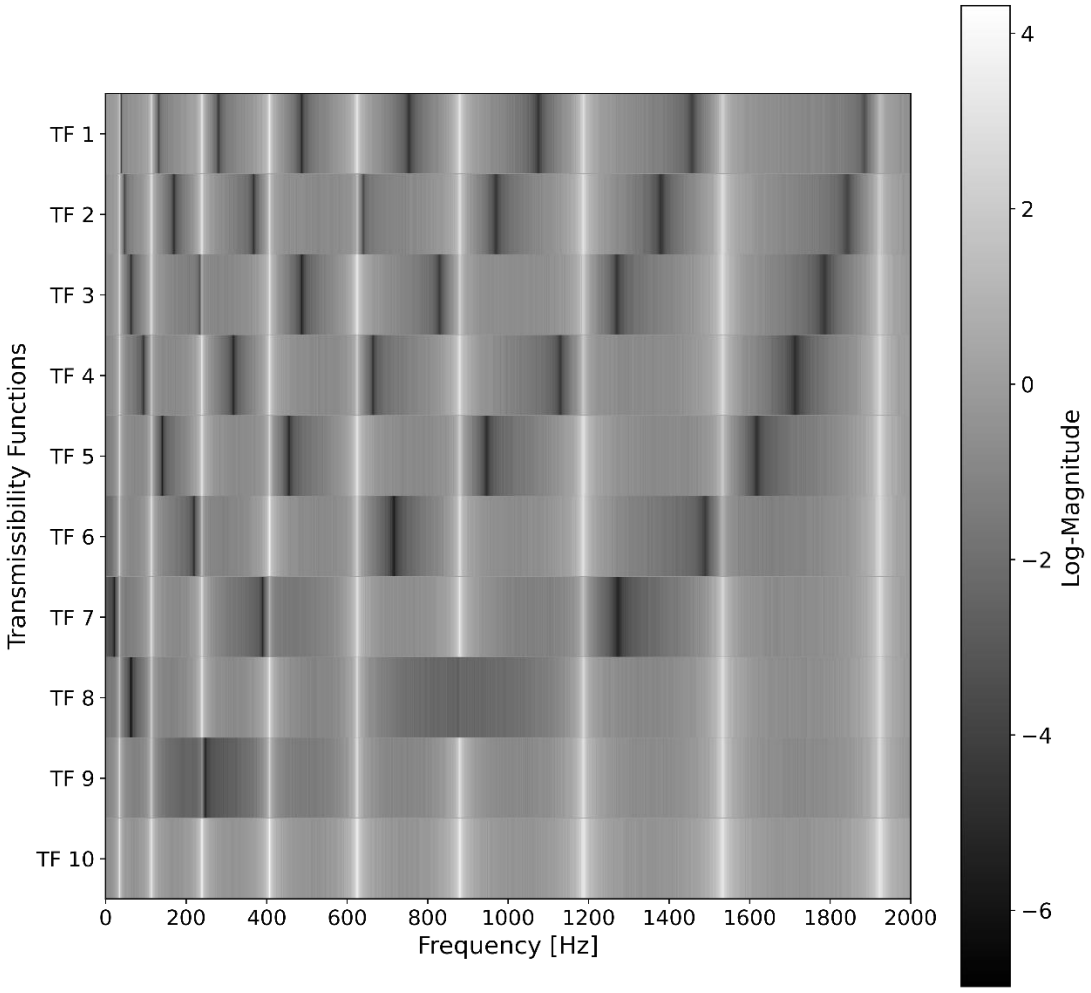

Figure 6 – Gray scale image representing the TFs for the damage scenario in which the element 12 presents $d\% = 12.5\%$ and noise is applied with $\delta = 30\%$.

The training and validation losses are shown in Figure 7, which show that the CNN has converged. Moreover, the optimal number of epochs, i.e., the number of epochs that lead to the best evaluation metric for the validation dataset, is identified to be 21.

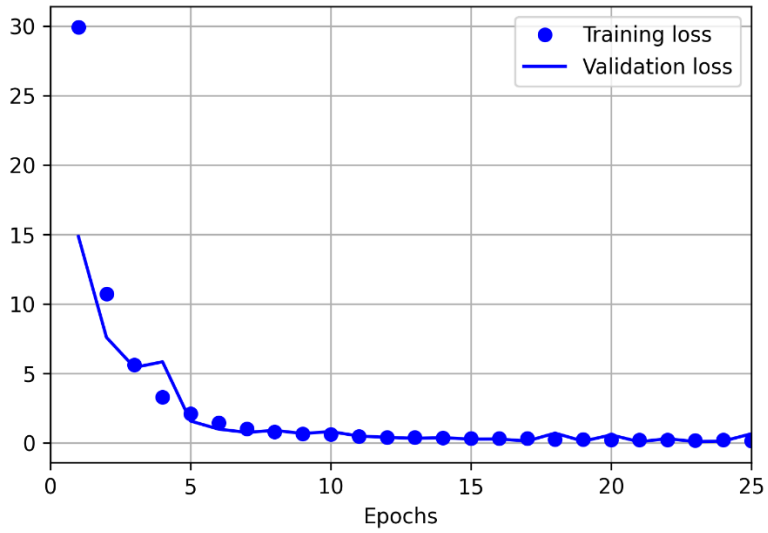

Figure 7 – Training and validation losses.

For the sake of brevity, even if a different CNN architecture is exploited, the interested reader is referred to [74] for more details about the CNN prediction capabilities related to the proposed case study. As an example, the CNN prediction for six damage scenarios are presented. In particular, the first damage scenario considers the element 12 with damage $d\% = 12.5\%$ and noise $\delta = 30\%$. The respective gray scale image is shown in Figure 6. In Figure 8 the absolute value of the predicted damage percentage for each one of the beam elements is reported. In this example, the CNN diagnosed the element 12 with $d^{CNN}\% = 12.64\%$, only 0.14 percentage points higher than the true damage condition. In addition to that, the CNN does not indicate any significant damage to other elements. In fact, the second largest $d_n^{CNN}\%$ value is the one with respect to element 20, in which $d^{CNN}\% = 0.98\%$.

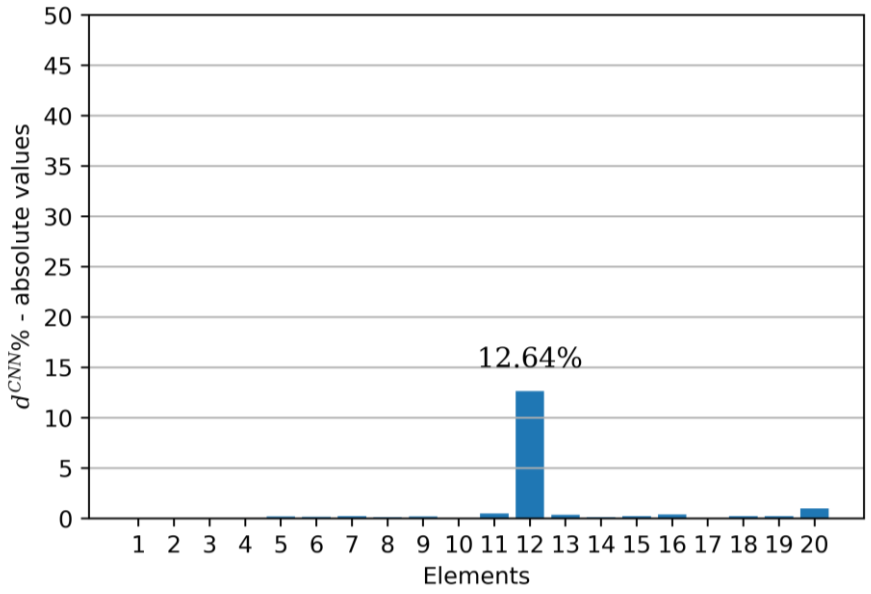

Figure 8 – CNN damage prediction for the damage scenario in which the element 12 is damaged with $d\% = 12.5\%$ and noise is applied with $\delta = 30\%$.

The second damage scenario considers element 7 with damage index $d\% = 22.5\%$ and the TFs corrupted with a noise given by $\delta = 30\%$. In particular, the predicted damage percentages are seen in Figure 9. In this scenario, the CNN predicted $d_7^{CNN}\% = 23.15\%$ and no other element is shown to be damaged, since $d_n^{CNN} < 0.5\%$ for all other elements.

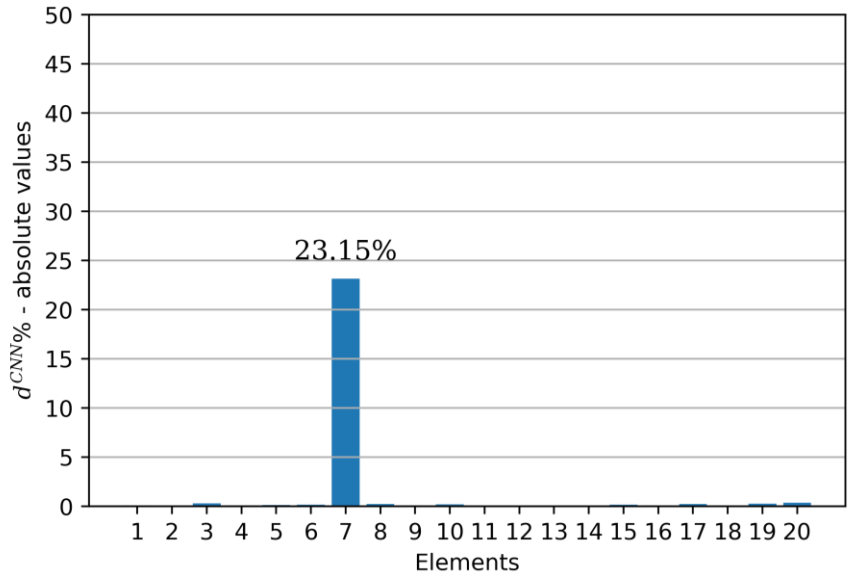

Figure 9 - Damage prediction for the damage scenario in which the element 7 is damaged with $d\% = 22.5\%$ and noise is applied with $\delta = 30\%$.

In Figure 10 it is possible to see the predicted damage values for the other four different damage scenarios, highlighting the CNN capabilities of predicting the damage at different conditions.

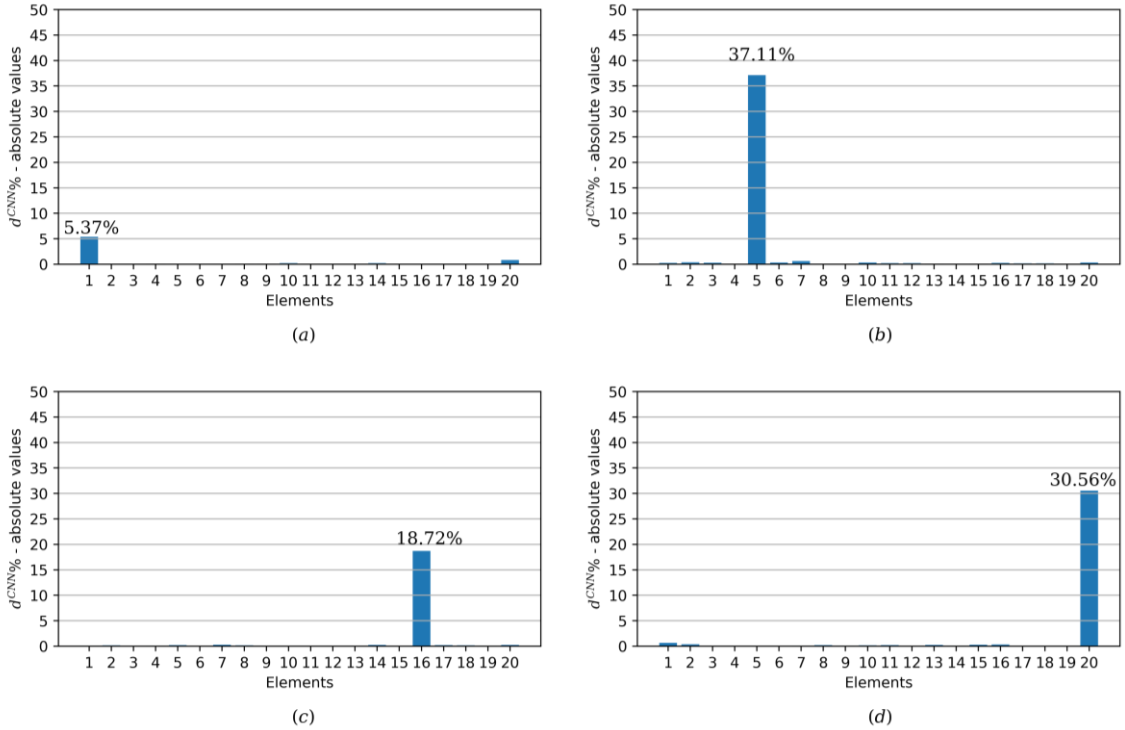

Figure 10 – CNN predictions for different damage scenarios. (a) Element 1 with $d\% = 7.5\%$ and $\delta\% = 30\%$(b) element 5 with $d\% = 37.5\%$ and $\delta\% = 10\%$; (c) element 16 with $d\% = 17.5\%$ and $\delta\% = 20\%$; (d) element 20 with $d\% = 27.5\%$ and $\delta\% = 30\%$.

5. XAI results: LRP algorithm

After analysing the damage diagnosis capabilities of the CNN, the LRP algorithm is exploited to interpret the CNN and understand how it is able to characterize damages. In order to do so, the iNNvestigate toolbox is utilized [75]. The Epsilon propagation rule is considered with $\epsilon = 3.0$. By applying the LRP algorithm, a relevance matrix R with dimensions 10x2000 which contains the relevance of each of the pixels of the gray scale input image is obtained, and, thus, the relevance for each one of the data points of the 10 considered TFs over the entire frequency window (i.e, 0-2000 Hz). Indeed, the relevance values contained in the first row of R are the relevances of TF 1, whereas the relevance values in the second row are related to TF 2, and so on. Such relevance matrix is turned into an expanded matrix R_{exp} of dimension 2000x2000 for visualization purposes, as done with matrix T. Finally, this expanded matrix can be plotted as a relevance heatmap. In Figure 11 a relevance heatmap is shown. In particular, it is originated for a damage scenario in which the element 12 is damaged by $d\% = 12.5\%$ and with a noise characterized by $\delta = 30\%$ applied. The CNN prediction for this scenario can be seen in the previous Section, in Figure 8.

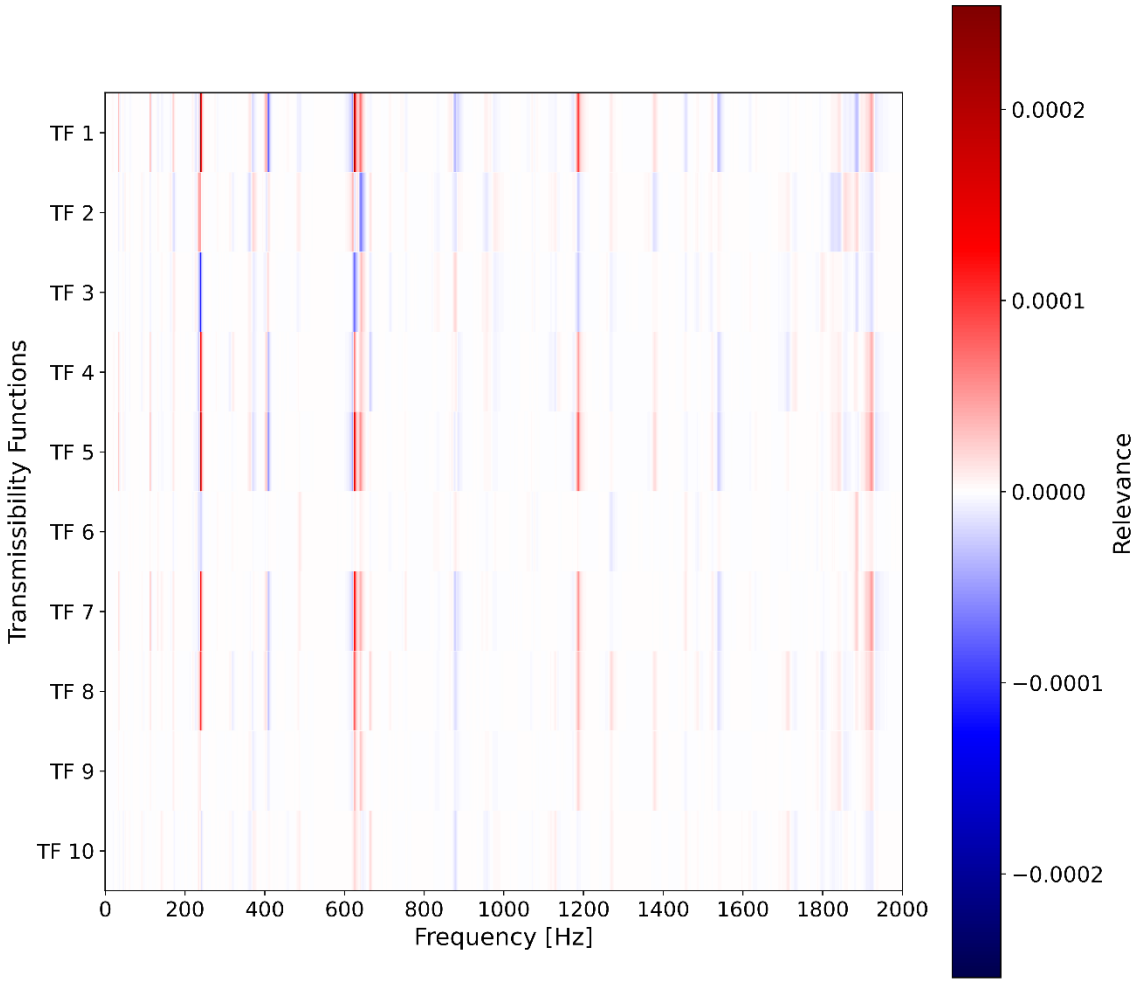

Figure 11 - Relevance heatmap for the damaged scenario in which the element 12 is damaged with $d\% = 12.5\%$ and noise is applied with $\delta = 30\%$.

By analysing the relevance heatmap, it is noticed that the relevance distribution is very sparse, i.e., some specific points concentrate the positive relevances. Indeed, most of the image is white, which stands for null relevance. It is also noticed that some specific frequency intervals appear to have higher relevance, even across multiple TFs. Moreover, all TFs present some positive relevance values, but in this case, these are more concentrated in TFs 1, 4, 5, 7 and 8. In order to better analyze the relevance values with the TFs features, the TFs are plotted with a colormap referring to their respective relevance value. In Figure 12 TFs 1, 4, 5 and 7 are reported. By analysing these plots, it is possible to verify that the peaks of the TFs are regions with positive relevance values. Indeed, as it was observed in [23], the peaks of the TFs are regions that present high sensibility to stiffness reduction, thus damaged elements. Therefore, it is reasonable to expect that the CNN would give them a higher relevance in order to perform the damage diagnosis.

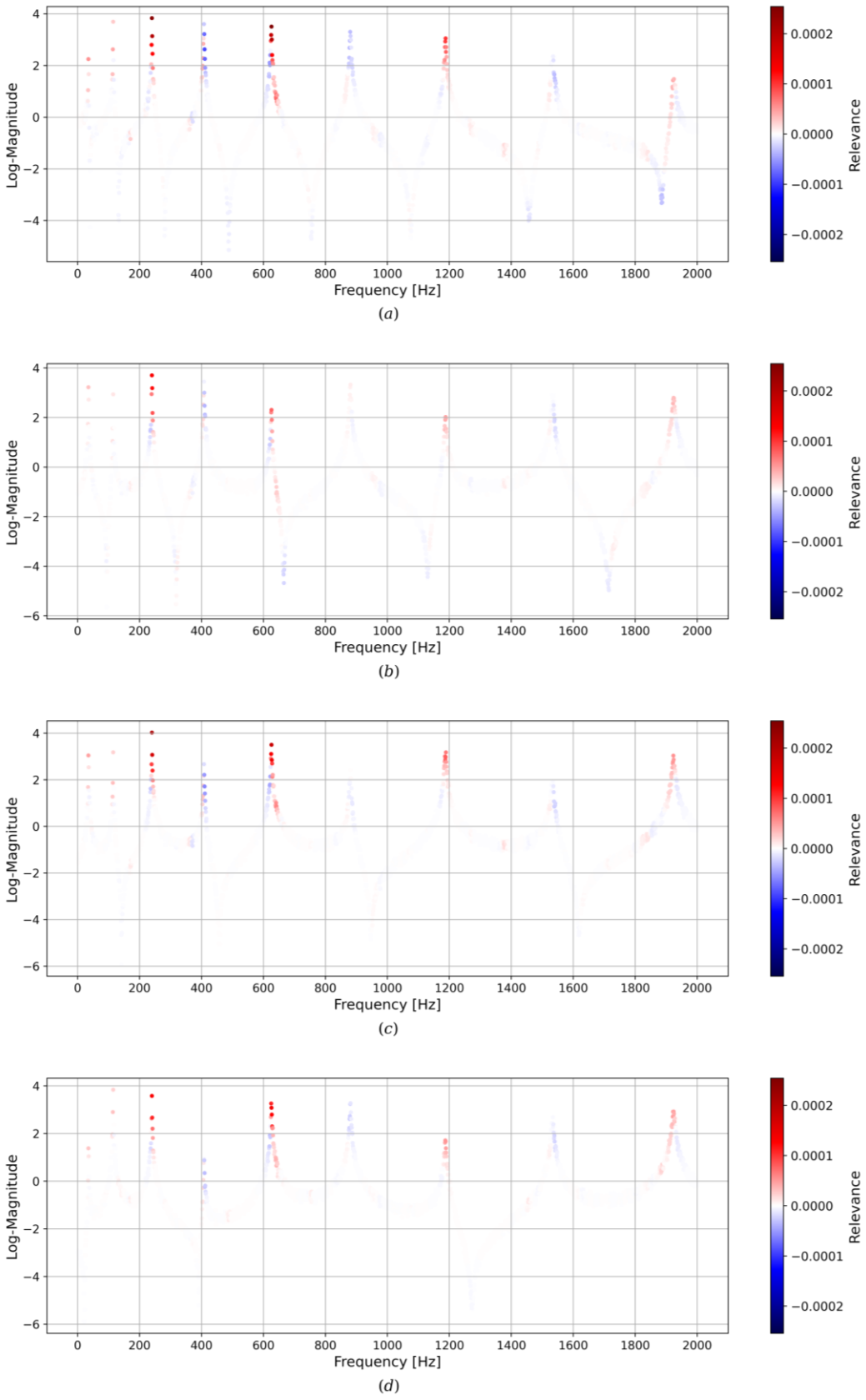

Figure 12 – TF 1 (a), TF 4 (b), TF 5 (c) and TF 7 (d) colored with respect to their corresponding relevance values for a damage scenario in which the element 12 is damaged with $d\% = 12.5\%$ and noise is applied with $\delta = 30\%$.

In order to further investigate the reasoning employed by the CNN, it is useful to introduce a damage index $DI_i(\omega)$, defined by the following Equation (17):

$$DI_i(\omega) = TF_d i - TF_h i \tag{17}$$

where $TF_d i$ refers the TF i evaluated for a generical damaged scenario, while $TF_h i$ denotes the same TF i but considering an undamaged structure (i.e., healthy scenario).

In particular, it is useful to investigate how it relates to the relevance values obtained for each TF. Therefore, considering the same scenario of Figure 9 in Section 4. (i.e., element 7 is damaged with $d\% = 22.5\%$ and noise is applied with $\delta = 30\%$.), in Figure 13 (a) TF 5 is shown, as well as its damage index, whereas, in Figure 14 (b) the TF 5 coloured in correspondence of its relevance values is reported and finally, in Figure 14 (c), the damage index computed for TF 5 also coloured in correspondence to its relevance values is shown.

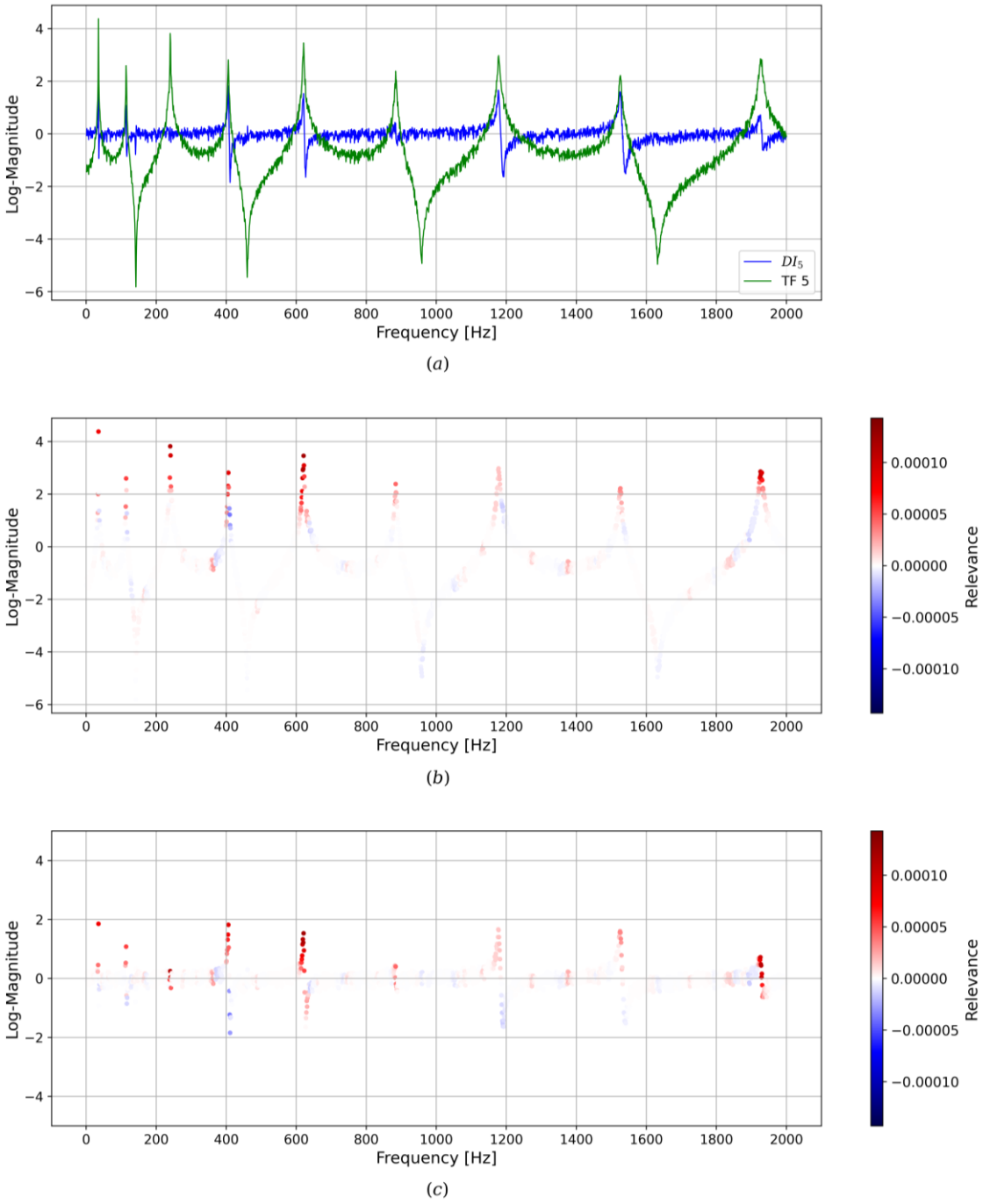

Figure 13 – (a) TF 5 and its respective damage index DI_5 considering the case in which the element 7 presents a damage $d\% = 22.5\%$ and with noise $\delta = 30\%$; (b) TF 5 colored with respect to its relevance values; (c) damage index computed for TF 5 colored with respect to the TF relevance values. Both considering the case in which the element 7 presents a damage $d\% = 22.5\%$ and with noise $\delta = 30\%$.

In particular, by analysing these three figures, the following points are observed: (i) the damage index achieves its highest values in correspondence of the TFs peaks. However, it is not every peak that present a significant damage index value. For instance, the peak at 883 Hz has a very small damage index when compared to other peaks, such as the one at 620 Hz; (ii) the points of highest

relevance values, which from now on will be referred as *relevance peaks*, occur mostly in correspondence of the TFs peaks, both the peaks that present high damage index value, e.g., the one near 620 Hz, and the ones that present low damage index value, e.g., the one at approximately 883 Hz; (iii) there are some relevance peaks, such as the one at 1378 Hz that are not in correspondence of any TF peak. These relevance peaks are classified as *spurious relevance peaks* and will be further investigated.

In order to understand why the CNN would give high relevance to a peak of a TF in a damaged scenario that does not present significant difference to its healthy counterpart, the damage index for other damaged scenarios is investigated. Considering the case in which the element 5 is damaged with $d\% = 22.5\%$ and a noise characterized by $\delta = 30\%$ applied, (CNN prediction $d_5^{CNN}\% = 22.4\%$), in Figure 14 (a) the TF 5 and its damage index are shown, in Figure 14 (b) the TF 5 colored in correspondence of its relevance values is reported and finally, in Figure 14 (c) the damage index computed for TF 5 also colored in correspondence to its relevance values is shown. It is possible to see that the majority of the relevance peaks are again in correspondence of the TFs peaks, but with some spurious relevance peaks also present. However, what should be noticed is that for this damaged scenario, the TF peak at frequency 865 Hz also has a relevance peak associated to it, but this time it presents a very high damage index. Therefore, for instance, the fact that, the TF peak localized at frequency 865 Hz does not change significantly for the damaged scenario with damaged element 7 but changes significantly for the damaged scenario with damaged element 5, could be used to locate the damaged element. Thus, for the case in which the element 7 is damaged, the TF peak at frequency 865 Hz is relevant for the CNN to predict that the damage is located at element 7 and not in element 5, for instance, justifying the relevance peak in correspondence of a TF peak with low damage index.

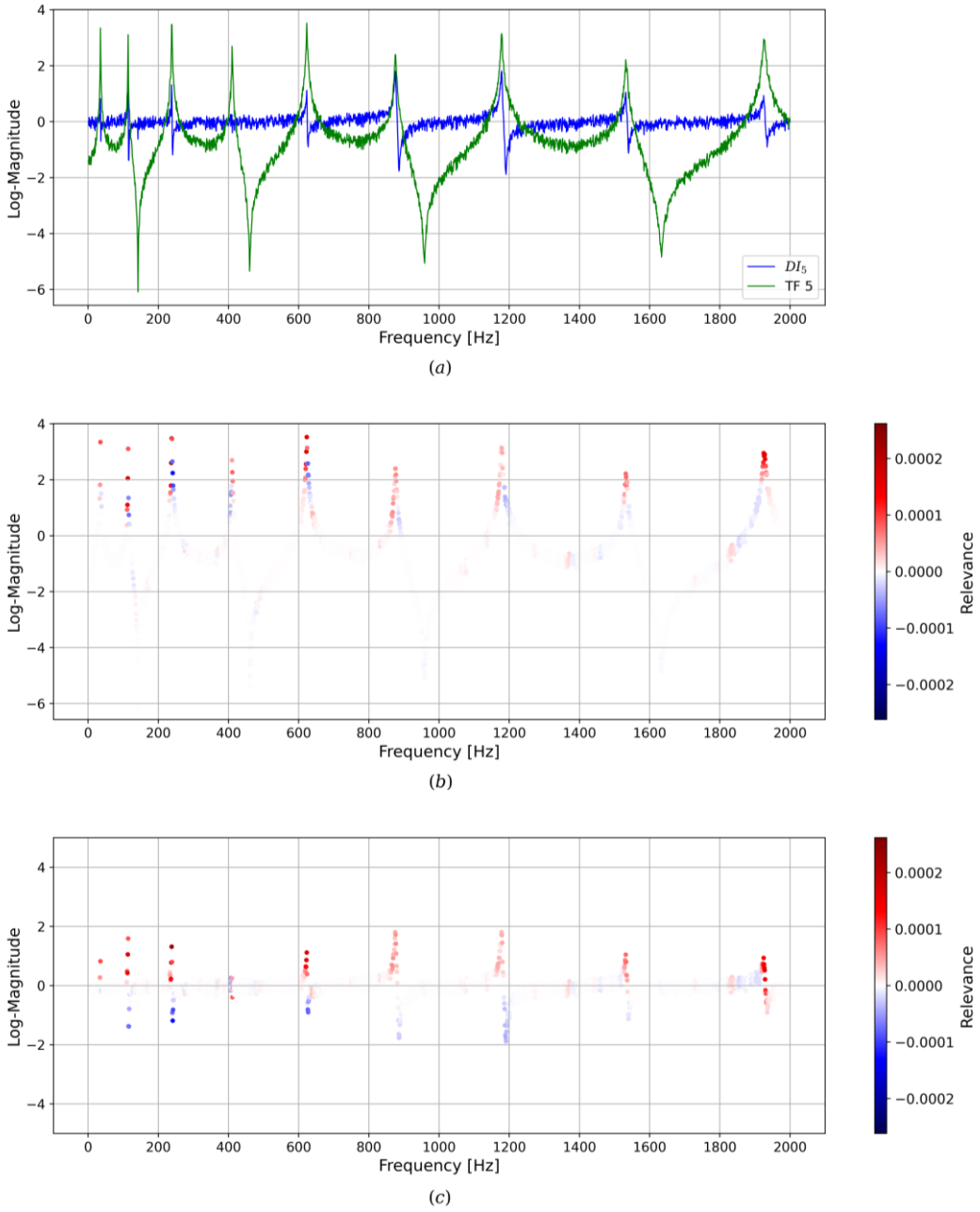

Figure 14 – (a) TF 5 and its respective damage index considering the case in which element 5 presents a damage $d\% = 22.5\%$ and with noise $\delta\% = 30\%$; (b) TF 5 colored with respect to its relevance values; (c) Damage index computed for TF 5 colored with respect to the TF relevance values. Both considering the case in which element 5 presents a damage $d\% = 22.5\%$ and with noise $\delta = 30\%$.

In order to investigate spurious relevance peaks, the plot of the TFs should be analyzed. In Figure 15 the TFs for the damaged scenario with the damaged element 7 that was previous discussed is shown. It is possible to see in the image that all TFs have peaks in the same frequencies. However, the TFs also present dips that are not common for all of them. In particular, in the frequency window between 1200 Hz and 1400 Hz, the TF 2 presents a dip at 1378 Hz, which is in the same frequency of

the spurious relevance peak previously identified. Therefore, it seems that the spurious peaks of TF 5 is related to the dips of TFs 2, 3 and 7. This relationship between different TFs could be justified by the kernel of the first convolutional layer of the CNN. Indeed, this kernel's dimension is (10,1), which indicates that it processes all TFs data together for each frequency.

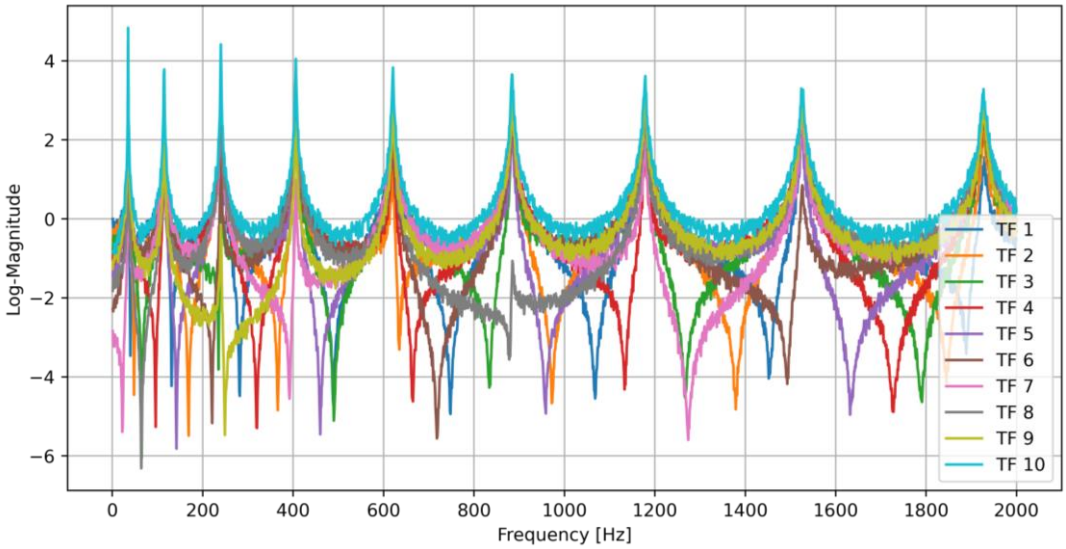

Figure 15 – Transmissibility functions for the case in which element 7 presents a damage $d\% = 22.5\%$ and with noise $\delta\% = 30\%$.

In order to confirm the relevance peaks and spurious relevance peaks patters previously observed, a statistical analysis is proposed. First, *correlated* relevance peaks are defined: a relevance peak correspondent to the i-th TF is considered a correlated peak if it occurs in a frequency window of width 18 Hz centred at any peak or dip of the i-th TF, otherwise it is considered a *spurious* relevance peak. A spurious relevance peak correspondent to the i-th TF is considered a *justified* spurious relevance peak if it occurs in a frequency window of width 18 Hz centred at any dip in any of the others j-th TFs ($i \neq j$). For instance, taking into consideration the damaged scenario investigated in Figure 13, the relevance peak of TF 5 at 620 Hz is a correlated relevance peak because TF 5 presents a peak at 620 Hz. However, the relevance peak at 1378 Hz is considered a spurious relevance peak, since there are no peaks from TF 5 that are close enough to it. Furthermore, this spurious peak is considered justified by the TF 2 dip at 1378 Hz.

The statistical analysis proposed consists in counting all relevance peaks and classifying them into correlated and spurious ones for different scenarios. Then, the spurious relevance peaks are divided between justified and non-justified ones. In order to perform this analysis, a dataset made of 1750 samples for each one of the damage elements (from element 1 to 20) is considered. For each of the damaged element, 7 groups of 250 samples corresponding to 7 different values of $d\%$ are considered, ranging from 7.5% to 37.5% with a 5% step. Each sample is corrupted by a numerical noise in the same way described for the training and testing datasets. In order to count the relevance peaks, the local maxima are compared with simple comparison with neighbouring values. However, for a local maxima to be considered a relevance peak, its height should be higher than a required height defined as $h_{req} = 0.1 * \max(R_{ij})\,;\ R_{ij} \in R$, i.e., for a given value, in order to be considered as a peak, it should be greater than 10% of the maximum relevance value in the relevance matrix R.

In Figure 16 it is plotted the normalized quantity of relevance peaks Q_p, given by Equation (18):

$$Q_p = \frac{N_p}{\max(N_p)} \tag{18}$$

where N_p is the sum of the number of relevance peaks of all the 1750 samples related to the damaged element position p. In addition, in Figure 17 the percentage of correlated and justified spurious peaks is reported. It is possible to see that percentage of correlated relevance peaks is always greater than 50%. Moreover, from the spurious peaks, more than 90% of them are classified as justified spurious peaks for all damage scenarios, except for the case in which the element 1 is damaged that the percentage of justified spurious peaks is lower than 90%. Thus, it is possible to confirm that for the observed scenarios, the relevance peaks are related to either TFs peaks or dips, showing that the CNN agrees with the physical intuition and by giving higher importance to the TF features that are highly damage sensible, as highlighted from the literature.

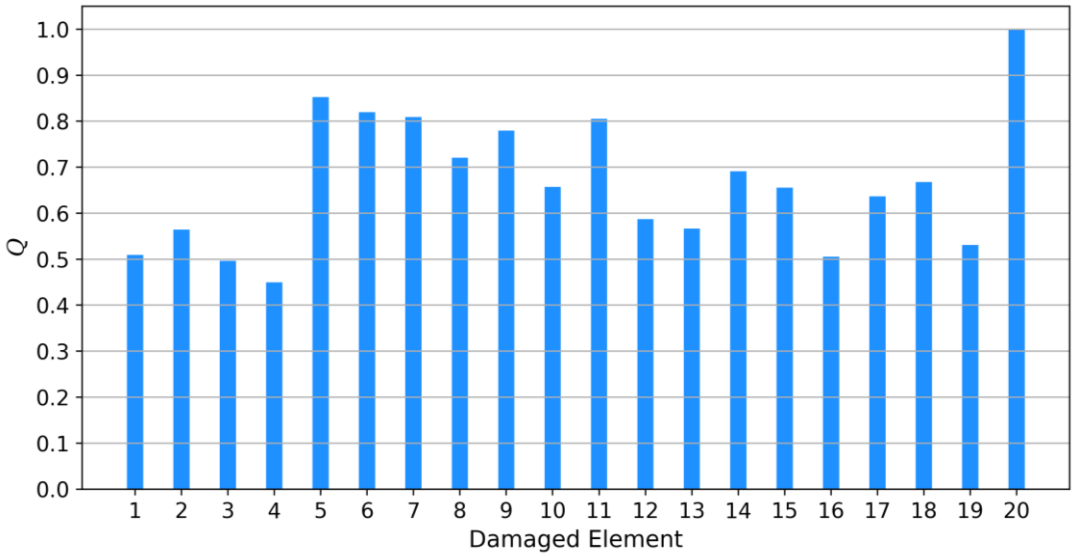

Figure 16 – Normalized quantity Q_p of relevance peaks for each damage scenario.

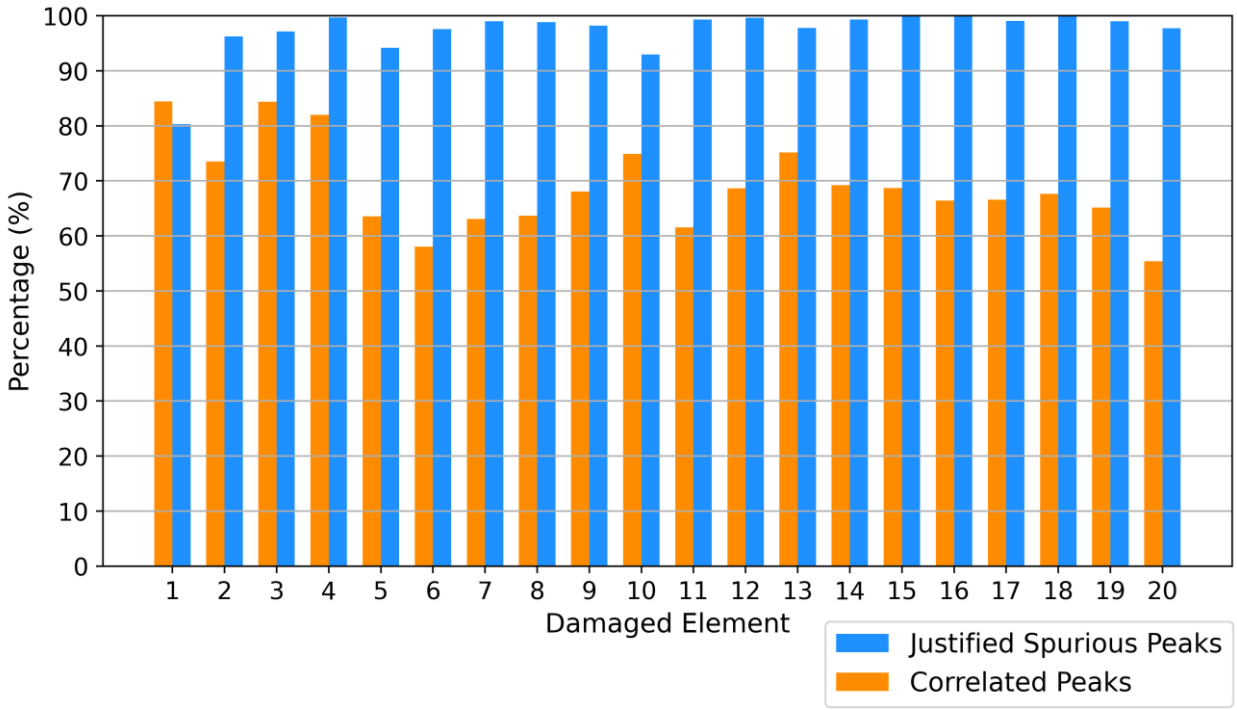

Figure 17 – Percentage of correlated peaks and justified spurious peaks for each damage scenario.

6. Conclusions

In order to increase the interpretability and, thus, the trustworthiness of a damage diagnosis algorithm based on a CNN which exploits transmissibility function (TF) data to perform damage detection, localization, and quantification in an output-only method, the layer-wise relevance propagation (LRP) algorithm was exploited to interpret the neural network. A numerical case study considering a structural beam was considered to implement the damage diagnosis method, which returned an output vector with the damage percentage of each element of the beam. Subsequently, the LRP was employed to obtain the relevance values for the input in order to understand the most important features for the CNN for a specific input-output pair. It was observed that the relevance values had some prominent values, i.e., relevance peaks, and that the majority of those was in correspondence of the TFs' peaks, therefore named correlated relevance peaks, which present a very high damage sensitivity. Moreover, it was noticed that from those that were not in correspondence of a TFs' peak, i.e., spurious relevance peaks, were justified by other TFs' dips. Such behavior is explained by the first convolutional layer, that processed all TFs together due to its kernel's dimensions. Therefore, it is identified that CNN is giving high importance to the most damage sensitive and important features of the TFs, which are their peaks and dips, agreeing with the existing physical knowledge. Finally, it is concluded that the proposed interpretability method was able to shed light on the reasoning behind the CNN exploited to perform the damage diagnosis task, enhancing its trustworthiness. For future works, the proposed explainability method could be employed in different non-simulated case studies.

For due purposes, I inform that the content of the work presented here will be fully presented at Escola Politécnica da Universidade de São Paulo as a final course work. Moreover, as a remark, I state that part of the work developed here was presented in the 8th World Conference on Structural Control and Monitoring in [76], which is yet to be published.

References

[1] H. M. Hashemian and W. C. Bean, "State-of-the-art predictive maintenance techniques," in *IEEE Transactions on Instrumentation and Measurement*, Oct. 2011, vol. 60, no. 10, pp. 3480–3492. doi: 10.1109/TIM.2009.2036347.

[2] C. Sbarufatti, A. Manes, and M. Giglio, "Application of sensor technologies for local and distributed structural health monitoring," *Struct Control Health Monit*, vol. 21, no. 7, pp. 1057–1083, 2014, doi: 10.1002/stc.1632.

[3] W. Fan and P. Qiao, "Vibration-based damage identification methods: A review and comparative study," *Structural Health Monitoring*, vol. 10, no. 1. pp. 83–111, Jan. 2011. doi: 10.1177/1475921710365419.

[4] Y. Yang, Y. Zhang, and X. Tan, "Review on vibration-based structural health monitoring techniques and technical codes," *Symmetry (Basel)*, vol. 13, no. 11, Nov. 2021, doi: 10.3390/sym13111998.

[5] C. R. Farrar, S. W. Doebling, and D. A. Nix, "Vibration-based structural damage identification," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 359, no. 1778. Royal Society, pp. 131–149, 2001. doi: 10.1098/rsta.2000.0717.

[6] N. M. M. Maia, J. M. M. Silva, E. A. M. Almas, and R. P. C. Sampaio, "Damage detection in structures: From mode shape to frequency response function methods," *Mech Syst Signal Process*, vol. 17, no. 3, pp. 489–498, 2003, doi: 10.1006/mssp.2002.1506.

[7] Z. Wang, R. M. Lin, and M. K. Lim, "Structural damage detection using measured FRF data," *Comput Methods Appl Mech Eng*, vol. 147, no. 1–2, pp. 187–197, 1997.

[8] Y. L. Xu, Q. Huang, S. Zhan, Z. Q. Su, and H. J. Liu, "FRF-based structural damage detection of controlled buildings with podium structures: Experimental investigation," *J Sound Vib*, vol. 333, pp. 2762–2775, 2014, doi: 10.1016/j.jsv.2014.02.010.

[9] S. Chesné and A. Deraemaeker, "Damage localization using transmissibility functions: A critical review," *Mech Syst Signal Process*, vol. 38, no. 2, pp. 569–584, 2013, doi: 10.1016/j.ymssp.2013.01.020ï.

[10] D. Verstraete, A. Ferrada, E. L. Droguett, V. Meruane, and M. Modarres, "Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings," *Shock and Vibration*, vol. 2017, 2017, doi: 10.1155/2017/5067651.

[11] S. Cofre-Martel, P. Kobrich, E. L. Droguett, and V. Meruane, "Deep Convolutional Neural Network-Based Structural Damage Localization and Quantification Using Transmissibility Data," *Shock and Vibration*, vol. 2019, 2019, doi: 10.1155/2019/9859281.

[12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.

[13] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, pp. 1–45, Jan. 2021, doi: 10.3390/e23010018.

[14] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/JPROC.2021.3060483.

[15] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-Wise Relevance Propagation: An Overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[16] Y. Bao, Z. Chen, S. Wei, Y. Xu, Z. Tang, and H. Li, "The State of the Art of Data Science and Engineering in Structural Health Monitoring," *Engineering*, vol. 5, no. 2, pp. 234–242, Apr. 2019, doi: 10.1016/j.eng.2018.11.027.

[17] M. Radzieński, M. Krawczuk, and M. Palacz, "Improvement of damage detection methods based on experimental modal parameters," *Mech Syst Signal Process*, vol. 25, no. 6, pp. 2169–2190, Aug. 2011, doi: 10.1016/j.ymssp.2011.01.007.

[18] E. Parloo, P. Guillaume, and M. van Overmeire, "Damage assessment using mode shape sensitivities," *Mech Syst Signal Process*, vol. 17, no. 3, pp. 499–518, 2003, doi: 10.1006/mssp.2001.1429.

[19] C. S. Hamey, W. Lestari, P. Qiao, and G. Song, "Experimental damage identification of carbon/epoxy composite beams using curvature mode shapes," *Struct Health Monit*, vol. 3, no. 4, pp. 333–353, 2004, doi: 10.1177/1475921704047502.

[20] T. J. Johnson and D. E. Adams, "Transmissibility as a Differential Indicator of Structural Damage," *J Vib Acoust*, vol. 124, no. 4, pp. 634–641, Sep. 2002, doi: 10.1115/1.1500744.

[21] Q. Chen, Y. W. Chan, and K. Worden, "Structural fault diagnosis and isolation using neural networks based on response-only data," *Comput Struct*, vol. 81, no. 22–23, pp. 2165–2172, Sep. 2003, doi: 10.1016/S0045-7949(03)00295-5.

[22] K. Worden, "Structural fault detection using a novelty measure," *J Sound Vib*, vol. 201, no. 1, pp. 85–101, 1997, doi: https://doi.org/10.1006/jsvi.1996.0747.

[23] K. Worden, G. Manson, and N. R. J. Fieller, "Damage detection using outlier analysis," *J Sound Vib*, vol. 229, no. 3, pp. 647–667, Jan. 2000, doi: 10.1006/jsvi.1999.2514.

[24] H. Zhang, M. J. Schulz, A. Naser, F. Ferguson, and P. F. Pai, "Structural Health Monitoring Using Transmittance Functions," *Mech Syst Signal Process*, vol. 13, no. 5, pp. 765–787, 1999, doi: https://doi.org/10.1006/mssp.1999.1228.

[25] K. Worden, G. Manson, and D. Allman, "Experimental validation of a structural health monitoring methodology: Part I. Novelty detection on a laboratory structure," *J Sound Vib*, vol. 259, no. 2, pp. 323–343, 2003, doi: 10.1006/jsvi.2002.5168.

[26] G. Manson, K. Worden, and D. Allman, "Experimental validation of a structural health monitoring methodology. Part II. Novelty detection on a Gnat aircraft," *J Sound Vib*, vol. 259, no. 2, pp. 345–363, 2003, doi: 10.1006/jsvi.2002.5167.

[27] G. Manson, K. Worden, and D. Allman, "Experimental validation of a structural health monitoring methodology: Part III. Damage location on an aircraft wing," *J Sound Vib*, vol. 259, no. 2, pp. 365–385, 2003, doi: 10.1006/jsvi.2002.5169.

[28] T. J. Johnson, R. L. Brown, D. E. Adams, and M. Schiefer, "Distributed structural health monitoring with a smart sensor array," *Mech Syst Signal Process*, vol. 18, no. 3, pp. 555–572, 2004, doi: 10.1016/S0888-3270(03)00002-5.

[29] A. Ghoshal, M. J. Sundaresan, M. J. Schulz, and P. F. Pai, "Structural health monitoring techniques for wind turbine blades," 2000.

[30] N. M. M. Maia, R. A. B. Almeida, A. P. V. Urgueira, and R. P. C. Sampaio, "Damage detection and quantification using transmissibility," *Mech Syst Signal Process*, vol. 25, no. 7, pp. 2475–2483, 2011, doi: 10.1016/j.ymssp.2011.04.002.

[31] H. R. Kess and D. E. Adams, "Investigation of operational and environmental variability effects on damage detection algorithms in a woven composite plate," *Mech Syst Signal Process*, vol. 21, no. 6, pp. 2394–2405, Aug. 2007, doi: 10.1016/j.ymssp.2006.11.010.

[32] D. H. Nguyen, T. T. Bui, G. de Roeck, and M. Abdel Wahab, "Damage detection in Ca-Non Bridge using transmissibility and artificial neural networks," *Structural Engineering and Mechanics*, vol. 71, no. 2, pp. 175–183, Jul. 2019, doi: 10.12989/sem.2019.71.2.175.

[33] W. Jacome, C. Yang, H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, *Deep Learning for NLP and Speech Recognition Related papers Recent Trends in Deep Learning Based Nat ural Language Processing "Deep Learning for Aspect-Based Sent iment …*, vol. 84. 2019.

[34] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[35] T. Mikolov, A. Deoras, D. Povey, L. Burget, and ˇ Cernock, "Strategies for Training Large Scale Neural Network Language Models," *IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 196–201, 2011.

[36] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process Mag*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun ACM*, vol. 60, no. 6, pp. 84–90, 2017, [Online]. Available: http://code.google.com/p/cuda-convnet/

[38] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 8, pp. 1915–1929, 2013, doi: 10.1109/TPAMI.2012.231.

[39] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A Survey of Deep Learning Techniques for Autonomous Driving," *J Field Robot*, vol. 37, no. 3, pp. 362–386, Oct. 2019, doi: 10.1002/rob.21918.

[40] G. Toh and J. Park, "Review of vibration-based structural health monitoring using deep learning," *Applied Sciences (Switzerland)*, vol. 10, no. 5. MDPI AG, Mar. 01, 2020. doi: 10.3390/app10051680.

[41] M. Azimi, A. D. Eslamlou, and G. Pekcan, "Data-driven structural health monitoring and damage detection through deep learning: State-ofthe- art review," *Sensors (Switzerland)*, vol. 20, no. 10, May 2020, doi: 10.3390/s20102778.

[42] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans Neural Netw Learn Syst*, pp. 1–21, Jun. 2021, doi: 10.1109/tnnls.2021.3084827.

[43] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*, Mar. 2018, vol. 2018-January, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[44] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[45] V. Ewald, R. M. Groves, and R. Benedictus, "DeepSHM: a deep learning approach for structural health monitoring based on guided Lamb wave technique," in *SPIE 10970, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, Mar. 2019, p. 19. doi: 10.1117/12.2506794.

[46] P. Seventekidis, D. Giagopoulos, A. Arailopoulos, and O. Markogiannaki, "Structural Health Monitoring using deep learning with optimal finite element model generated data," *Mech Syst Signal Process*, vol. 145, Nov. 2020, doi: 10.1016/j.ymssp.2020.106972.

[47] L. Rosafalco, M. Torzoni, A. Manzoni, S. Mariani, and A. Corigliano, "Online structural health monitoring by model order reduction and deep learning algorithms," *Comput Struct*, vol. 255, Oct. 2021, doi: 10.1016/j.compstruc.2021.106604.

[48] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1. MDPI AG, pp. 1–45, Jan. 01, 2021. doi: 10.3390/e23010018.

[49] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Mag*, vol. 40, pp. 44–58, 2019.

[50] A. Hanif, X. Zhang, and S. Wood, "A Survey on Explainable Artificial Intelligence Techniques and Challenges," in *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW*, 2021, pp. 81–89. doi: 10.1109/EDOCW52865.2021.00036.

[51] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in neural information processing systems*, 2017, vol. 30. [Online]. Available: https://github.com/slundberg/shap

[52] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accid Anal Prev*, vol. 136, Mar. 2020, doi: 10.1016/j.aap.2019.105405.

[53] S. Chehreh Chelgani, H. Nasiri, and M. Alidokht, "Interpretable modeling of metallurgical responses for an industrial coal column flotation circuit by XGBoost and SHAP-A 'conscious-lab' development," *Int J Min Sci Technol*, vol. 31, no. 6, pp. 1135–1144, Nov. 2021, doi: 10.1016/j.ijmst.2021.10.006.

[54] K. Zhang, P. Xu, and J. Zhang, "Explainable AI in deep reinforcement learning models: A SHAP method applied in power system emergency control," in *2020 IEEE 4th Conference on Energy Internet and Energy System Integration: Connecting the Grids Towards a Low-Carbon High-Efficiency Energy System, EI2 2020*, Oct. 2020, pp. 711–716. doi: 10.1109/EI250167.2020.9347147.

[55] M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M. J. Segovia-Vargas, "Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending," *IEEE Access*, vol. 8, pp. 64873–64890, 2020, doi: 10.1109/ACCESS.2020.2984412.

[56] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in Fintech Risk Management," *Front Artif Intell*, vol. 3, Apr. 2020, doi: 10.3389/frai.2020.00026.

[57] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[58] M. Kuzlu, U. Cali, V. Sharma, and Ö. Güler, "Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools," *IEEE Access*, vol. 8, pp. 187814–187823, 2020, doi: 10.1109/ACCESS.2020.3031477.

[59] Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics*, vol. 12, no. 2. MDPI, Feb. 01, 2022. doi: 10.3390/diagnostics12020237.

[60] A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," *Front Artif Intell*, vol. 4, Sep. 2021, doi: 10.3389/frai.2021.752558.

[61] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, "Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation," *arXiv preprint arXiv:1611.08191*, Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.08191

[62] I. Ullah, A. Rios, V. Gala, and S. McKeever, "Explaining deep learning models for tabular data using layer-wise relevance propagation," *Applied Sciences (Switzerland)*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/app12010136.

[63] Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching, "Explaining therapy predictions with layer-wise relevance propagation in neural networks," in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, Jul. 2018, pp. 152–162. doi: 10.1109/ICHI.2018.00025.

[64] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Front Aging Neurosci*, vol. 10, no. JUL, 2019, doi: 10.3389/fnagi.2019.00194.

[65] F. Eitel *et al.*, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation," *Neuroimage Clin*, vol. 24, Jan. 2019, doi: 10.1016/j.nicl.2019.102003.

[66] W. Yan *et al.*, "Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6. doi: 10.1109/MLSP.2017.8168179.

[67] H. Cho, E. K. Lee, and I. S. Choi, "Layer-wise relevance propagation of InteractionNet explains protein–ligand interactions at the atom level," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-78169-6.

[68] H. Bharadhwaj, "Layer-Wise Relevance Propagation for Explainable Deep Learning Based Speech Recognition," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2018, pp. 168–174. doi: 10.1109/ISSPIT.2018.8642691.

[69] J. Grezmak, J. Zhang, P. Wang, K. A. Loparo, and R. X. Gao, "Interpretable Convolutional Neural Network through Layer-wise Relevance Propagation for Machine Fault Diagnosis," *IEEE Sens J*, vol. 20, no. 6, pp. 3172–3181, Mar. 2020, doi: 10.1109/JSEN.2019.2958787.

[70] O. Mey and D. Neufeld, "Explainable AI Algorithms for Vibration Data-based Fault Detection: Use Case-adadpted Methods and Critical Evaluation," *arXiv preprint arXiv:2207.10732*, Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.10732

[71] T. Kumagai, M. Kohiyama, and T. Yamashita, "Interpretation of Deep Neural Network for Damage Pattern Classification Using Phase Plane," in *Proceedings of The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications*, 2020.

[72] L. Lomazzi, S. Fabiano, M. Parziale, M. Giglio, and F. Cadini, "On the explainability of convolutional neural networks processing ultrasonic guided waves for damage diagnosis," *Mech Syst Signal Process*, vol. 183, p. 109642, Jan. 2023, doi: 10.1016/j.ymssp.2022.109642.

[73] E. M. Dogo, O. J. Afolabi, N. I. Nwulu, B. Twala, and C. O. Aigbavboa, "A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 92–99. doi: 10.1109/CTEMS.2018.8769211.

[74] M. Parziale, L. Lomazzi, M. Giglio, and F. Cadini, "Vibration-based SHM exploiting a combination of CNN and autoencoders for temperature effects neutralization," *Struct Control Health Monit*, 2022.

[75] M. Alber *et al.*, "iNNvestigate Neural Networks!," 2019. [Online]. Available: http://jmlr.org/papers/v20/18-540.html.

[76] M. Parziale, P. H. Silva, M. Giglio, and F. Cadini, "[TO BE PUBLISHED] Explainable convolutional neural networks for SHM based on transmissibility functions," *8th World Conference on Structural Control and Monitoring (8WCSCM)*, vol. To be published.

Abstract in lingua italiana

Una vasta gamma di settori, come l'ingegneria civile, meccanica e aerospaziale, ha impiegato unità di monitoraggio, o *structural health monitoring* (SHM), per consentire un funzionamento più sicuro ed efficiente dei dispositivi e strutture. Queste unità si basano su algoritmi per elaborare i dati acquisiti dalla struttura del sistema per eseguire il rilevamento, la localizzazione e/o la quantificazione dei danni in tempo reale. Per fare ciò, tra le diverse tipologie di segnali acquisibiliquelli vibrazionali sono stati ampiamente e con successo impiegati, poiché alcune proprietà ,come le frequenze naturali, lo smorzamento modale e le forme modali di un sistema, dipendono dalle sue proprietà strutturali, che possono essere soggette a cambiamenti indotti dai danni. In particolare, le funzioni di trasmissibilità, o *transmissibility functions* (TFs), hanno suscitato molto interesse poiché non dipendono dal modulo della forzante di input (solo dal suo punto di applicazione), semplificando quindi fortemente il processo di acquisizione e quindi poi di diagnosi. Recentemente, questo tipo di dati strutturali utilizzati dai sistemi SHM sono diventati sempre più accessibili grazie ai significativi progressi tecnologici e alla riduzione dei costi dei sensori, creando così un ambiente molto favorevole per l'applicazione di modelli di *deep learning* nel campo SHM. Sebbene questi modelli abbiano mostrato risultati molto promettenti in termini di accuratezza della previsione, questi di solito hanno un costo in termini di interpretabilità del modello a causa della loro crescente complessità. In effetti, è più difficile fidarsi di algoritmi che mancano di interpretabilità, che è una caratteristica fondamentale nelle applicazioni ingegneristiche pratiche come i sistemi SHM. Al fine di aumentare l'interpretabilità di tali modelli, sono stati proposti molti metodi di intelligenza artificiale spiegabile, o *Explainable AI* (XAI), come l'algoritmo di *layer-wise relevance propagation* (LRP). Dunque, in questo lavoro, una *convolutional neural network* (CNN), un tipo specifico di modello di *deep learning*, è stata utilizzata per elaborare delle TFs al fine di eseguire il rilevamento, la localizzazione e la quantificazione dei danni, e successivamente interpretata attraverso l'uso dell'algoritmo LRP. Considerando un caso studio numerico di una trave strutturale con diversi scenari di danno, i valori di rilevanza restituiti dall'algoritmo XAI sono stati indagati attraverso un'analisi statistica. Si è osservato che la maggior parte delle caratteristiche più rilevanti per la CNN sono quelle in cui le TFs risultano essere più sensibili al danno, in accordo con quanto evidenziato dalla lettura e, quindi, dalle conoscenze fisiche riguardanti le TFs.

Keywords: Explainable AI, Structural Health Monitoring, Convolutional neural networks, Transmissibility Functions, Layer-wise Relevance Propagation