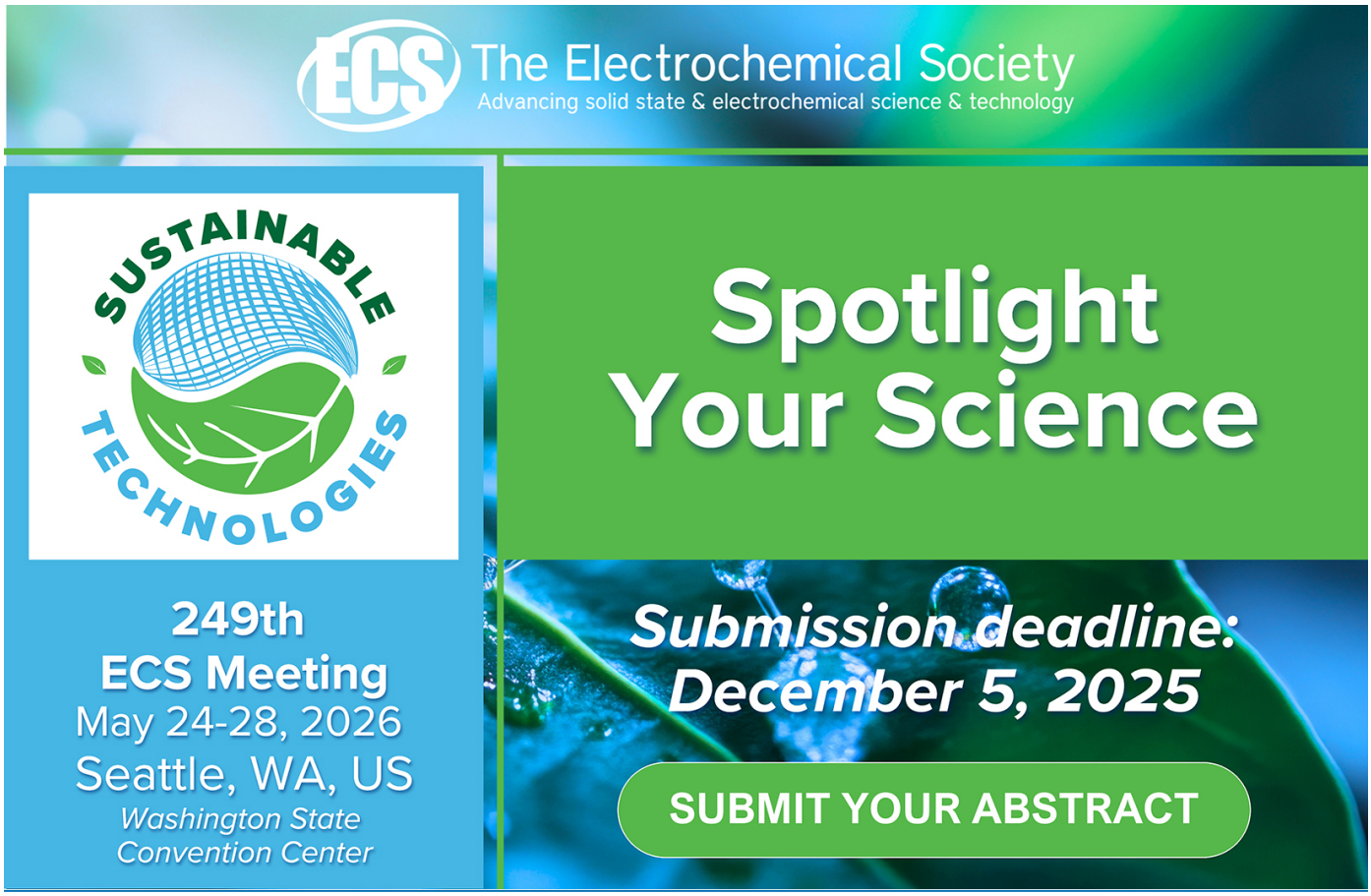
Smart Materials and Structures

PURPOSE-LED PUBLISHING™

A unified CNN approach for guided wave-based damage detection, damage size estimation and reliability assessment demonstrated on a complex composite structure



OPEN ACCESS
IOP Publishing
Smart Materials and Structures
Smart Mater. Struct. **33** (2024) 105034 (12pp)
https://doi.org/10.1088/1361-665X/ad6aba

A unified CNN approach for guided wave-based damage detection, damage size estimation and reliability assessment demonstrated on a complex composite structure

Oliver Schackmann[1], **Vittorio Memmolo**[1,2,*] **and Jochen Moll**[1]

[1] Department of Physics, Goethe University Frankfurt, Max-von-Laue-Straße 1, 60438 Frankfurt am Main, Germany
[2] Department of Industrial Engineering, Universitá degli Studi di Napoli 'Federico II', Via Claudio 21, 80125 Naples, Italy

E-mail: vittorio.memmolo@unina.it

Received 19 September 2023, revised 30 July 2024
Accepted for publication 2 August 2024
Published 23 September 2024

Abstract

This work presents a novel unified Convolutional Neural Network approach where broadband ultrasonic guided waves signals are processed in such a way that damage is first detected (binary classification) and then its severity assessed on continuous scale (multi-class classification) without resorting to different procedures. To test the learning approach and assess the classification procedures, a hyperparameter optimization is first carried out to determine the best data processing procedure. Then, the performance of the network is evaluated thoroughly. The results demonstrated the relationship between the model's performance and SHM system parameters, including excitation signal, pre-processing approach and the number of paths utilized within a sparse distributed transducer network. Furthermore, the damage location is an important influence factor. In addition to that, ensemble voting is demonstrated to be the most accurate approach to achieve high reliability in damage detection and size assessment. The results show the capability of the proposed methodology (i) to detect early damage with highest possible accuracy (ii) to estimate the dimension of damage with limited error and reasonable accuracy, and (iii) to assess the reliability of the whole monitoring system through damage size estimation combined with a critical damage size approach.

Keywords: structural health monitoring, damage detection, ultrasound, aerospace structures, Lamb waves, machine learning

* Author to whom any correspondence should be addressed.

1. Introduction

Structural Health Monitoring (SHM) of composite structures is a promising yet challenging approach to achieve lighter and safer structures, in line with the ambitious target of climate neutrality in many transportation fields [1]. Among several techniques available, guided ultrasonic waves (GUWs) have raised the attention of the scientific community because they are sensitive to hidden defects and can propagate over long distances with limited attenuation [2]. The current GUW-based damage identification strategies rely on identifying changes in physical parameters of propagating waves to discriminate as to whether there exist damaged states or not. Despite many approaches already available in the literature [3], main challenges are still present and related to: (i) the number of sensors adopted to achieve accurate diagnosis, which can have a dramatic impact on the benefits achievable when integrating SHM in vehicle structural health management [4], and (ii) how to reduce the loss of diagnostic information using meaningful features retrieved from recorded signals. These concerns raise steeply when attempting to fully characterize damage (occurrence, location and severity) and to quantify and qualify the SHM system [5]. Within this scenario, artificial intelligence promises to have a great impact. Because of the superior capabilities of machine learning approaches in recognizing and classifying available patterns in a dataset, they have demonstrated a significant improvement in traditional damage identification algorithms [6]. There are several examples in the literature where machine learning approaches are adopted successfully for damage prediction [7, 8]. In addition, big efforts have been carried out to connect ubiquitous sensing and big data processing of critical information to achieve a digital structural health monitoring pipeline and achieve near real-time and online damage assessment [9]. However, many issues still arise in terms of explainability and interpretability as well as metrics definition. The latter is still mostly related to the loss of information when attempting to achieve less intense data streaming (e.g. using damage features). Indeed, most methods rely on signal processing to get scalar information from the GUW signals, which inevitably reduce diagnostic information without any knowledge about loss severity and consequent reduced accuracy [10]. Indeed, reliability has a crucial impact on the whole decision-making process because inspection and maintenance is taken according to the location, severity and possible propagation of the identified damage.

In this context, Convolutional Neural Networks (CNNs) offer the advantage of processing multidimensional data directly, as opposed to relying on extracted scalar information. This offers the possibility to avoid unknown information reduction. However, it is still very difficult to achieve data-efficient training as they usually require a large amount of data, either making model based training more appropriate (requiring a further effort for transfer learning to the actual scenario) [11] or requiring data dimensionality reduction through indices [12]. In addition, there is still a lack of applications in order to achieve a unified approach for structural damage detection, localization and quantification [5]. Furthermore, damage position and/or severity are usually roughly estimated through binary or multi-label classification, while regression should be employed to pave the way towards quantification [13], which is required for system qualification. In this regard, explaining the reasoning process is not trivial, requiring additional effort in clear data processing and definition of accuracy metrics.

In this view, the present paper is intended to present a novel CNN based approach combining binary and multi-class predictors. The latter turns into a regression approach through the inherent structure of the neural network architecture. This provides continuous rather than discrete information about damage size and is a fundamental step towards reliability assessment. The main results lay the groundwork for a unified machine learning-based damage diagnosis. A use case with relatively limited experimental data for CNN training is exploited to: (i) test the capabilities of the proposed machine learning-based framework with limited sources, (ii) bring evidence of the accuracy of the algorithms involved to fully characterize the damage, and (iii) demonstrate the reliability of the whole monitoring system in critical damage size estimation.

The remaining of the paper is organized as follows. The methodology is first discussed, including details of the overall concept, the experimental setup description and the approach used for data analysis. Afterwards, results are reported showing first the hyperparameter optimization and then the main results in terms of influencing parameters along with damage position effect. Finally, accuracy results are reported showing the way to find a good compromise between accuracy definition and acceptable error. Concluding remarks including main achievements and future research plans close the paper.

2. Methodology

2.1. Conceptual description of the approach

In this work, we consider defects as structural abnormalities that occur during manufacturing and that can be detected using conventional non-destructive testing approaches, such as ultrasonic testing or radiography. On the other hand, a damage emerges during regular operation of the structure, such as a fatigue damage. The new idea in this work is to feed CNNs with GUW measurements for the binary classification of damage (hit/miss) and its size prediction (increasing severity) while paving the way towards reliability assessment thereof. In this way a unified approach for damage characterization is inherently achieved.

Due to their structure, CNNs are particularly suited to process images. To this end, GUW signals are converted into a suitable (image-like) form following the steps shown in figure 1, showing the data processing pipeline for binary and multi-class classification. The measurements are sequentially elaborated to input a unified CNN model for damage detection and severity assessment.

After sourcing and preprocessing chirp signals, the time history is transformed into a spectrogram (SP) using

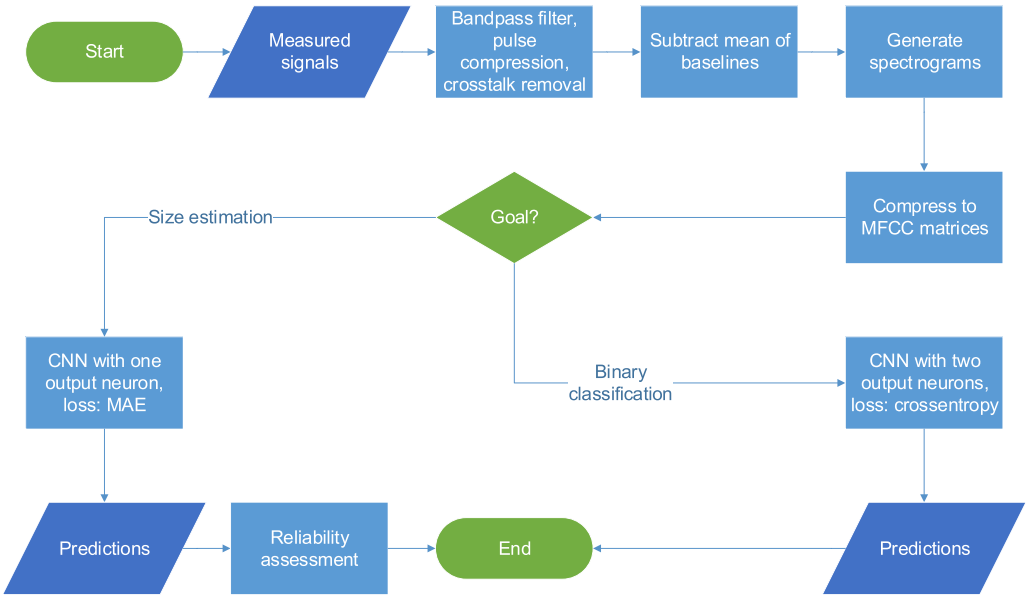


Figure 1. Algorithmic flowchart for binary classification and damage size estimation.

short-time Fourier transform. The latter has the advantage of preserving most of frequency and time information and representing a suitable input for ML algorithm. Nonetheless, to enable a data-efficient training phase, the use of smaller images is preferable. Therefore, the signals are further compressed into Mel-Frequency Cepstrum Coefficients (MFCC) matrices and then used for two purposes (each with its own neural network). The former is exploited to investigate the ability to discriminate between undamaged and damaged status. Afterwards, the latter model is used to predict the damage size. Two slightly different networks are used for the two intended tasks. For both models, a CNN presented in [14] successfully applied to the Fashion-MNIST dataset served as a template. This has been designed as a slightly more sophisticated version of the MNIST dataset, which is very well known for testing machine learning algorithms, and consists of 70 000 grayscale images of size 28 $\times$ 28 of fashion items from ten categories. Since the size of these images is very similar to those of the MFCC matrices (20 $\times$ 32), the same model architecture should work well for the task at hand.

It is worth noting that the process shown in figure 1 ends with the detection of damage (binary classification) and the size assessment (multi-class classification). Despite this begs the question about reliability and paves the way to further actions, any health management procedure is not discussed hereinafter, because it needs the reliability output inputting a specific risk assessment procedure.

2.2. Description of the experimental setup

The experimental data set used for the AI-based data analysis is available at the Open Guided Waves (OGW) online platform and has been acquired using a carbon fiber reinforced polymer (CFRP) specimen equipped with a network of piezoelectric transducers [15]. An omega stringer is attached to the plate as typically designed for aircraft components. The sketch and a photo of the specimen with stringer is shown in figure 2.

Twelve piezoceramic transducers are applied to the plate alongside the stringer on both sides and excited in pitch catch mode by a linear chirp signal from 20 kHz to 500 kHz The advantage of using this diagnostic signal with respect to a single frequency pulse is to collect a broadband response of the structure in one shot with a dramatic reduction of measurement time [16, 17]. Using chirp signal, Lamb waves are excited in a broader range, but still in the form of A_0 and S_0 modes. The effect of this choice allows to exploit different modes which have different sensitivity to damage, keeping most of the information. The multimodal coexistence on the detection accuracy can improve detection as evaluated through hyperparameter optimization in the present investigation. The piezoceramic transducers are used to excite the signal while a Handyscope HS5 (TiePie Engineering) is employed to generate and record the signals via digital-to-analog and analog-to-digital conversion (14 bit resolution). A broadband amplifier PD200 (PiezoDrive Ltd Shortland, NSW 2307, Australia) is used to boost the signals and then feed it to a dedicated multiplexing device used to actuate and also measure all actuator-receiver pairs by time-division multiplexing [18]. Further details about data acquisition and characterization of wave propagation in this specific specimen is reported in [15].

To simulate the effect of damage on the guided wave propagation, artificial reversible damage of multiple sizes is conceived with metallic plates, which are attached to the specimen using a vacuum sealant tape [15]. The representative damage is necessary to evaluate SHM systems in a standardized way. A detailed justification of this approach can be found in [19]. The reference scatterer should be able to represent the key features of the damage (size, interaction with the excited

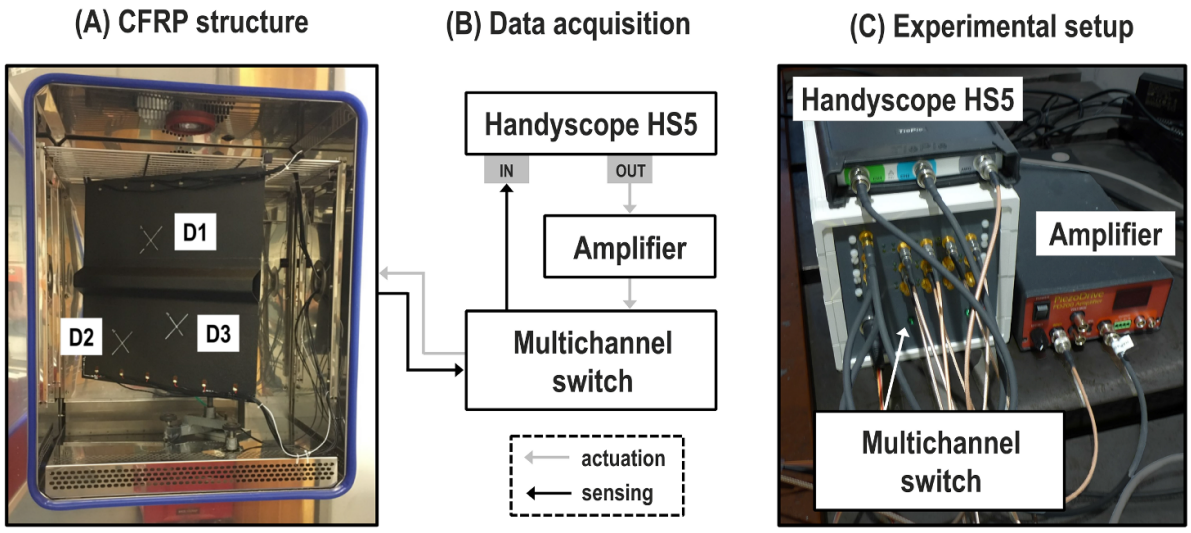


Figure 2. (A) Photo of the CFRP specimen in the climatic chamber to realize controlled atmospheric conditions in terms of constant temperature and relative humidity (including the three damage positions D1–D3). (B) Workflow of the data acquisition procedure during the experimental campaign (C) Photo of the key components during data acquisition. Further information regarding the experimental setup can be found in [15].

Table 1. Size of damage applied in different state conditions.

State ID	Baseline	#1	#2	#3	#4	#5	#6
Damage size (mm^2)	0	49.48	111.80	198.75	310.55	447.19	608.68
State ID	#7	#8	#9	#10	#11	#12	#13
Damage size (mm^2)	671.07	795.52	1006.2	1242.23	1503.12	1788.85	2090.53

GW-modes).For this purpose, an elliptically shaped damage was chosen. Its size is motivated by the projected surface of real impacts [20]. A total of 13 damage sizes ranging from 49.48 mm^2(damage #1) up to 2090.53 mm^2 (damage #13) are used according to table 1. During the acquisition of ultrasonic data, each damage is attached on 3 defined positions, as depicted in figure 2. Baseline signals (pristine conditions) are taken before and after each damage is placed on a defined position, with a total of twenty acquisitions. The structured process of data acquisition consisted of seven phases:

- Phase 1 Five baseline measurements of the pristine structure are recorded.
- Phase 2 Reference defects are attached to the plate at position D1 and measurements are carried out.
- Phase 3 Baseline measurements are recorded without any applied reference damage.
- Phase 4 Reference defects are attached to the plate at position D2 and measurements are carried out.
- Phase 5 Another five baseline measurements were recorded without any applied reference damage.
- Phase 6 Reference defects are attached to the plate at position D3 and measurements are carried out.
- Phase 7 Five more baseline measurements are acquired without any applied reference damage.

The specimen is placed in a climate chamber and temperature is controlled showing a very low standard deviation temperature gradient between different regions of the plate. For each of the 13 damage sizes and every damage position, the measurement was independently repeated 5 times, resulting in 5 statistically independent measurements. Each time the procedure includes the following steps: fixation of the model defect with tacky tape, measurement, detaching, removal of tacky tape, renewed fixation.

It is possible to implement the designed approach with only one representative transducer pair (on opposite sides of the stringer) using the data analysis method described in the following section. However, better results are achieved with a generalization to a distributed transducer network, which is discussed in section 3.

2.3. Data analysis methods

As shown in figure 1, data processing can be split into different stages. After labelling the data, signal preprocessing is carried out mostly to clean data. At this stage, the chirp signal is first processed with band pass and matched filters to enhance signal quality. The first one is a Butterworth filter of first order with band gates fixed at 10 kHz to 510 kHz. The matched filter is then applied pointing to the actuation signal. The resulting waveform is windowed to exclude crosstalk, which is electromagnetic coupling arising in the field and at the acquisition board. This procedure is essential to increase signal to noise ratio as well as to reduce misleading information due to measurements uncertainties. Then the data transformation stage

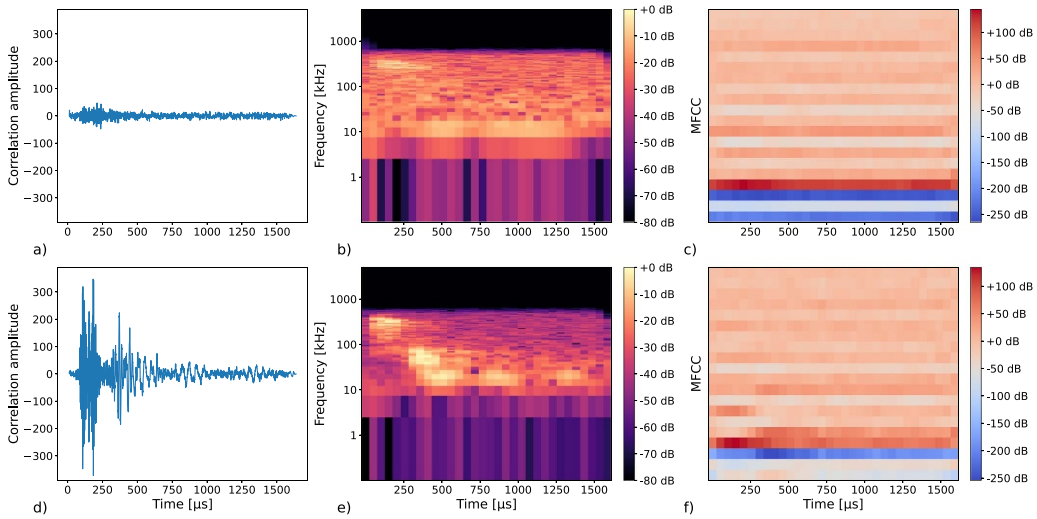


Figure 3. Signal processing with baseline data displayed in the top row and data of damage #13 in the bottom row. (a), (d): Differential signals (b), (e): Spectrograms of the differential signals (c), (f): MFCC matrices of the differential signals.

starts with the calculation of the differential signal. It consists in subtracting the average of the 20 preprocessed baselines (also referred to as super-baseline) from all signals acquired on the pristine and damaged conditions. The idea is to find any differences between each signal and one reference condition (super-baseline). This provides the actual datasets for analysis. These are further edited to get first to the SP representation, which is suitable input to feed CNN, and then to MFCC, which compress the data (see section 2.1). For these steps, the Python library librosa [21] is used, keeping default settings and adjusting the sampling frequency. Figure 3 shows the main results of signal processing applied to one signal among baseline dataset (first row) and one signal among damage #13 dataset (second row). The time signals in (a) and (d) represent the differential signal obtained by subtracting the super-baseline as mentioned above. This comparison allows to track any change due to wave-damage interaction in a wide-band spectrum. As highlighted in the figure, results in pristine and damaged condition are pretty different. Pictures in (b) and (e) represent the corresponding SP of the time signals. Instead pictures in (c) and (f) represent the corresponding MFFC of the time signals.

To build the learning procedure up, the obtained MFCC matrices are divided into three different groups. Data from two out of the three damage positions, along with 80% of the baseline signals, are used for training and validation with an internal split of 75% for the former step and 25% for the latter step, respectively. Instead, the testing set consists of measurements from the remaining damage position and the rest of the baselines. In this way, the CNN model first learns from the training set how to distinguish among damage and undamaged state. By determining different metrics on the validation set, it can be assessed whether the CNN can handle unseen data from known positions. Having the test set data out of these allows to generalize the procedure and assess the more realistic scenario of encountering unknown measurements at unknown damage positions.

Afterwards, the classification is established with two different network architectures, as explained in figure 1. For the binary classification, Géron's network [14] is slightly modified. This is achieved by experimenting and implementing generally known recommendations such as the use of He-weight initialization for layers with ReLU activation functions. The last layer consists of only two neurons, corresponding to the number of cases to be distinguished.

While binary classification is providing an acceptable means of compliance for qualification, in view of continuous damage detection it is worth achieving severity assessment, estimating the risk of overcoming a critical damage size (usually set up during the engineering design phase). In addition, having a continuous rather than discrete severity prediction enables the quantification of the proposed approach (e.g. minimum detectable damage) and the following qualification of the system. To predict the size of the damage, only small changes in the network structure are necessary. The main difference from Géron's network [14] consists in establishing a continuous value for the prediction instead of the conventional discrete class label. *Therefore, the classification task turns into a regression problem.* Accordingly, the last layer now contains only a single neuron without an activation function.

In both classification approaches, a batch size of 32 and the Adam optimizer with cosine decay as a learning rate schedule are used. In the binary prediction, the loss function categorical cross-entropy is to be minimized within 100 epochs. Instead, in the multi-class prediction, the mean absolute error is to be minimized within 300 epochs. The discrepancy in training duration stems from the complexity of the task under consideration. Overall, the CNN architecture is described in table 2, which includes parameters of each single layer from data input (bottom) to prediction output (top).

Table 2. CNN architecture. The parameters of each single layer are described from data input (bottom) to prediction output (top).

Layer	Filters	Size (Out)	Filt. Size	Stride	Padding	Act. Func.
Fully con.	—	2/1	—	—	—	Softmax/—
Dropout	—	64	—	—	—	—
Fully con.	—	64	—	—	—	ReLU
Dropout	—	128	—	—	—	—
Fully con.	—	128	—	—	—	ReLU
Flatten	—	2048	—	—	—	—
Average Pooling	256	2×4	2×2	2	valid	—
Convolution	256	5×8	3×3	1	same	ReLU
Average Pooling	128	5×8	2×2	2	valid	—
Convolution	128	10×16	3×3	1	same	ReLU
Average Pooling	64	10×16	2×2	2	valid	—
Convolution	64	20×32	3×3	1	same	ReLU
Input	—	20×32	—	—	—	—

3. Results and discussions

Although one transducer pair is enough to detect the presence of damage (see section 3.2), a sensor network is necessary to monitor a wide area. The approach can be extended to a cluster of transducers whenever they are used to monitor the same structural area and the same damage occurrence.

The estimation procedure becomes more complicated with the increasing number of paths available and the dependencies that arise among them in relation to the damage location. In addition, the training time increases in proportion to the greater amount of information available. The following sections examine how the results depend on the number of paths. The cases considered are that all, only from opposite sides, only from the same side or only a single path in the middle of the plate ($T_3 - T_9$) is used.

3.1. Hyperparameter optimization

Using the Python library Optuna [22], the network hyperparameters which may affect the prediction are analyzed. This allows to automate hyperparameter search and generate 100 combinations, which are tested for regression and classification, respectively. Its TPESampler (Tree-structured Parzen Estimator) makes the optimization efficient while its MedianPruner stops unpromising trials before the completion of all epochs to save time. In this step, the data is divided into just two parts. The data from damage positions D2, D3 and the first 80% of the baselines form the training set, while the validation set consists of the remaining data. This is where the accuracy is to be maximized and the mean absolute error minimized.

The hyperparameters considered are: frequency content (chirp vs. 40 kHz toneburst), use of signal differentiation (True vs. False), considered path channels (all, opposite, same side), standardized inputs (True vs. False), learning rate (between 10^{-3} and 10^{-5}), and learning rate decay alpha (between 10^{-1} and 10^{-6}). The values of the hyperparameters leading to the best results are shown in table 3. Both classification and regression have in common that the employment of chirp signals, the subtraction of baselines as a step in the preprocessing and the use of paths running across the plate work best. However, the standardization of the MFCC matrices is preferred for classification, while this step is better omitted for regression.

Table 3. The best parameters found with Optuna. For binary classification the accuracy is maximized, for damage size estimation the mean absolute error is minimized.

Parameter	Classification	Regression
Frequency	chirp	chirp
Differential	True	True
Channels	opposite	opposite
Standardization	True	False
Learning Rate	1.23×10^{-4}	7.71×10^{-5}
Alpha	1.0×10^{-1}	1.0×10^{-3}
Value	99.7 %	262 mm^2

The best combination is then adopted as the starting point to vary one hyperparameter alone, while the others are held constant. By varying a single hyperparameter, results can be generated for different training-validation-test combinations. This makes it possible to determine more precisely which hyperparameters have a major influence and to what extent the location of the damage affects the results.

3.2. Influence of individual parameters and damage positions

The results obtained for the networks with the best hyperparameters and with individual varied hyperparameters with and without ensemble voting are displayed in figures 4–7, respectively. In the classification, the dashed black line indicates a threshold below which the network provides no added value. This is due to the fact that the test set consists of significantly more measurements for damaged plates than undamaged ones. Therefore, a network that always predicts 'damage' would achieve an accuracy of 94.2%. The size and color of the dots indicate the location of the damage in the test set.

Below each of the results for the test sets the predictions of a voting ensemble are presented. So far, the network takes

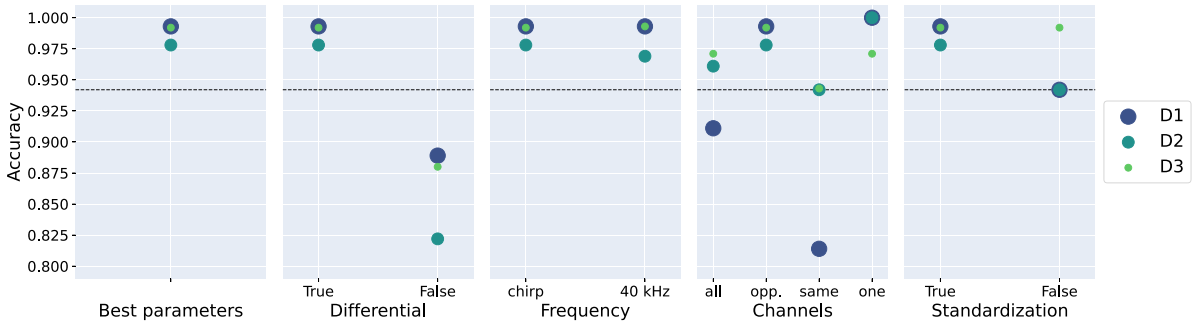


Figure 4. Results for binary classification on the test set with variation of individual parameters.

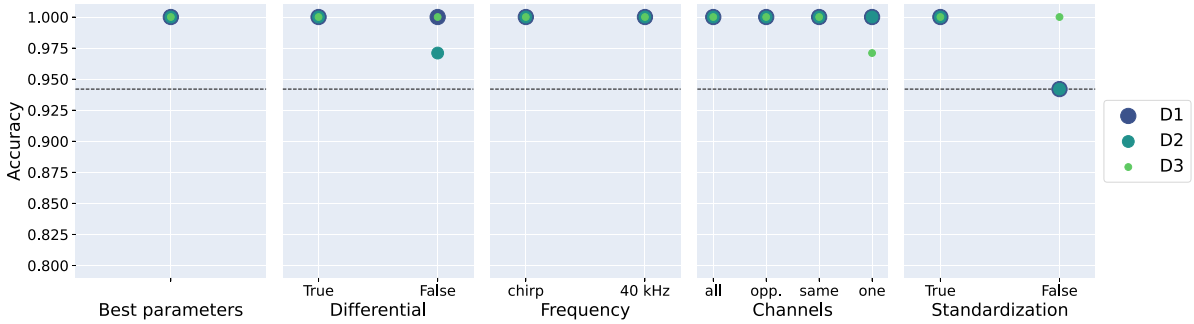


Figure 5. Results after ensemble voting for binary classification on the test set with variation of individual parameters.

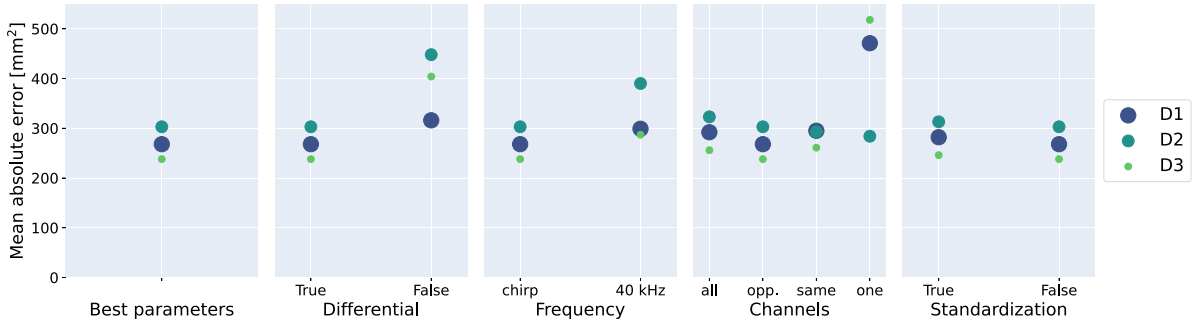


Figure 6. Results for damage size estimation on the test set with variation of individual parameters.

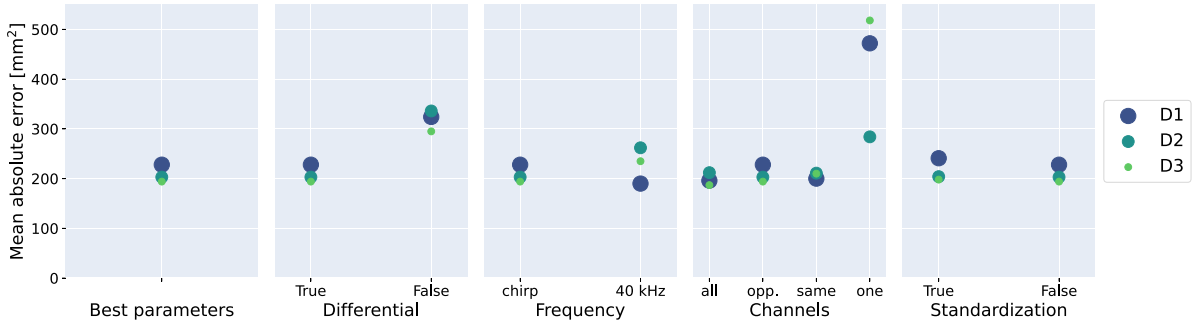


Figure 7. Results after ensemble voting for damage size estimation on the test set with variation of individual parameters.

an MFCC matrix based on a single measurement and a single channel as input and predicts the probabilities for the class labels or the estimated damage size. However, these predictions can be significantly improved by averaging the predictions of many variations of the same measurement. Each of these variations corresponds to a different path of the GUW across the measurement object.

For the binary classification, the probabilities of both classes are averaged and the higher mean probability is selected as the prediction of the ensemble. For the regression

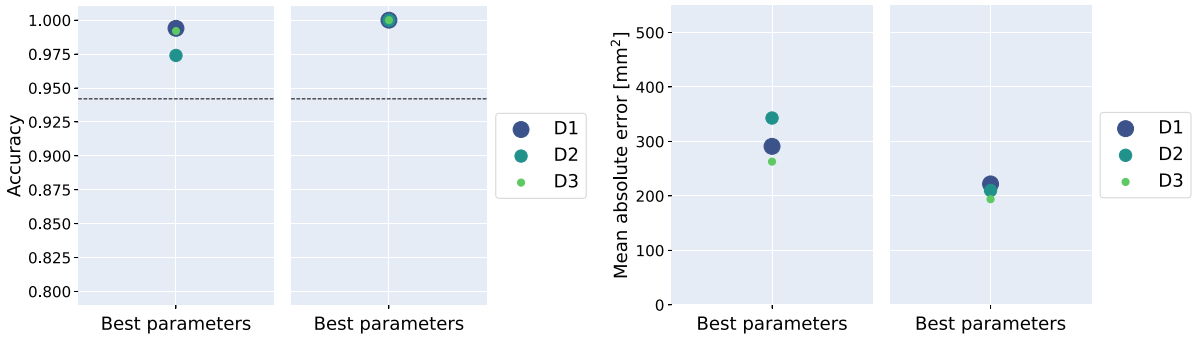


Figure 8. Results with EfficientNetV2B0 on the test set with variation of individual parameters. For the binary classification and the regression, the values before the ensemble voting are shown on the left and after it on the right.

problem, the mean value is used as the final estimate. This procedure therefore increases the accuracy as it reduces the variability in the individual predictions.

In the case of damage detection, it can be observed that only by using the difference signal, the opposite channels and by standardizing the matrices, a performance above the self-set threshold can be achieved for all damage locations. With other hyperparameters, the network is not able to reliably distinguish between undamaged and damaged plates. In some cases, it is even worse on this test set than a hypothetical network that would predict 'damage' in every case. Using ensemble voting, however, it is possible to achieve an accuracy of 100% in most cases, even with non-optimal parameters. Only the standardization must not be omitted.

The results for estimating the damage size are qualitatively similar. Again, the most significant differences can be seen when using the difference signal. Interestingly, it is not always the same parameters that produce the best results both with and without ensemble voting. For example, before ensemble voting, the use of the paths across the plate is recommended, but after ensemble voting the value for all channels is slightly better. It is possible that these small discrepancies have only random causes such as different weight initializations and could disappear after averaging several training runs. This could be determined in future work with more (computing) resources.

It can be observed that even a single channel delivers perfect results for the classification for two of the three positions. However, it should be noted that no ensemble voting is possible here that could improve the prediction further for the third position. When estimating the damage size, on the other hand, the differences to the predictions of many paths are much greater. For this task in particular, it is clearly recommended to use many paths in order to obtain a useful estimate.

In general, it can be seen that the position of the damage has an influence on the result, but is usually not decisive for the general effectiveness of the method. For most of the hyperparameter combinations tested, damage at position D3 is the easiest to detect and the most accurate to estimate in size, while for D2 the results are often the worst. In the latter case, one explanation is that this damage location is furthest away from the signal sources and changes in the waves that have already attenuated at this point are more difficult to detect.

Finally, it is worth noting that dispersion and multimodal characteristics of guided waves are generally beneficial in damage detection as different modes have different sensitivity to damage. As a consequence, keeping most of the information using chirp signals can help improving detection. Indeed, the influence of multimodal coexistence on the detection accuracy is evaluated through hyperparameter optimization comparing chirp excitation with the 40 kHz toneburst excitation, which generally gives the best performance for this kind of problem [23]. They have a very comparable sensitivity to damage in terms of accuracy. However, broadband excitation is less affected by the influence factors (including damage position). In addition, it does not depend upon the frequency (or Lamb wave mode) sensitivity to a specific damage type. Accordingly, it results in the most general approach that can be applied to SHM system.

A final attempt to achieve even better results is made by implementing a larger CNN. EfficientNetV2B0 [24] is used here, which is known for high accuracy in image recognition tasks with impressive efficiency. The number of its parameters is approximately 6 million, which is about 10 times larger than the previously used network. This also means that the training time is around 10 times longer. However, this does not pay off in this case, as can be seen in figure 8. The metrics for the training and validation data are significantly better for the EfficientNet than for the self-designed network, but the results on the test data are almost identical. This suggests that the larger network tends to overfit and memorize the training data. As a result, it works excellently on known positions, but for signals from unknown positions it does not provide any added value compared to the more compact model. One possible explanation for the lack of improvement in performance is that EfficientNet was created for larger images and more extensive training data sets. For our task, it therefore does not seem to be a solution to simply utilize networks with larger numbers of parameters.

3.3. Prediction results

Another compilation of the results of the own network with the best parameters for estimating the damage size can be seen in

Table 4. Results using several metrics for the training, validation and test set. The mean and standard deviation for the results from the three damage locations are reported.

	MAE (mm^2)	a_{NN}	a_{fix}	$a_{10\%}$
Training	180 ± 2	0.39 ± 0.01	0.67 ± 0.01	0.20 ± 0.01
Validation	194 ± 7	0.36 ± 0.01	0.64 ± 0.02	0.17 ± 0.01
Test	270 ± 28	0.26 ± 0.03	0.52 ± 0.04	0.14 ± 0.02
Test ensemble	209 ± 15	0.30 ± 0.10	0.62 ± 0.04	0.16 ± 0.03

table 4. In addition to the MAE, other metrics are listed that examine the performance of the model on various aspects:

- Mean absolute error (MAE). To compare forecasts with the eventual outcomes, the error of all predictions is measured with respect to the actual value and averaged.
- Accuracy nearest neighbors (a_{NN}). Each prediction is rounded to the closest value from the set of available discrete damage sizes. In this way, a confusion matrix can also be created directly from the regression task. The accuracy obtained corresponds to the proportion of points located on the diagonal.
- Accuracy fixed (a_{fix}). A prediction is classified as correct if it falls within an interval of specified size. The idea is that the prediction error within this interval is neglected.
- Accuracy percentage ($a_{10\%}$). All predictions with relative deviation smaller than 10% from the true damage size are counted as correct.

Based on these interpretation approaches, the prediction results are reported in figure 9. The first plot (a) shows scattered prediction data and regression thereof. The latter models the predicted value and can be compared with perfect prediction to give an idea of accuracy of the model response. In addition, the 95% confidence (green bars) is defined for each predicted value starting from MAE estimation. Rounding the prediction value to the closest class (identified) it is possible to generate the confusion matrix from the regression prediction (b), from which a_{NN} comes out. The plots in (c) and (d) are created after applying ensemble voting, but are otherwise based on the same principles. In addition to the smaller fluctuations compared to the plots above, it is noticeable that the model tends to estimate values that are too low, particularly for large damages. This could be explained by the fact that there is an upper limit for the damage size in the training data. The predictions consequently rarely exceed this value and therefore tend to be too low, especially after ensemble voting.

Figure 10 shows how a_{fix} and $a_{10\%}$ can be adopted to bound the prediction and define the region of correct prediction constraining constant or percentage error, respectively. In other words, starting from the mechanics of composites and damage tolerance, the accuracy can be computed by giving meaning to an acceptable error. Knowing this acceptable error threshold and given a critical damage size (set to 671 mm^2 for this application [15]), it becomes possible to estimate a predicted value that signals the presence of this critical damage according to the accuracy given in table 4. As can also be seen in plot (a), predictions of exactly 0 mm^2 would therefore be necessary for the baseline. As a consequence, even the smallest deviation is considered incorrect. A similar strict rule applies for small damage sizes, likely returning incorrect predictions even with negligible errors. All this results in an apparently low accuracy term $a_{10\%}$ despite the good prediction performance. Otherwise, the fixed error is able to contain more data and provides much better accuracy at the cost of earlier warning (441 mm^2 instead of 574 mm^2). Finally, figure 10(b) proposes a totally different approach based on the 95% confidence bound of the prediction model. This interpretation relies on solid statistical assumptions because the 95% confidence interval is commonly interpreted as there is a 95% probability that the true linear regression line of the population will lie within the confidence interval of the regression line calculated from the sample data. Hence, the signal response warning the presence of critical damage can be defined as from the lower confidence bound. This approach returns later warning (641 mm^2) with statistically meaningful prediction.

4. Concluding remarks

This paper looks into a novel unified GUW-based SHM approach based on CNN to detect and classify damage in composite structures. Starting from the GUW raw data recorded by ultrasonic transducers permanently installed on a stiffened composite plate, different post-processing techniques are adopted to feed a CNN based on SP and MFCC. This allows to reduce the number of variables strongly, with a consequent reduction of data needed for the training and validation process. In detail, the novel CNN based approach combines binary and multi-class predictors for damage assessment. While the former can with the latter one turned into a regression approach through the inherent structure of the neural network architecture. This is effectively used to perform a continuous estimation of the damage size.

The main theoretical achievement relies in the use of a specific CNN architecture for multi-class classification which we have conceived to have a unified approach for different damage classification stages. Indeed, it is based on a single neuron for the final layer, which returns in a regression analysis and, consequently, in a continuous damage assessment with high accuracy when ensemble voting is adopted

Results show the best combination of hyperparameters for the CNN as well as the importance of each single parameter in

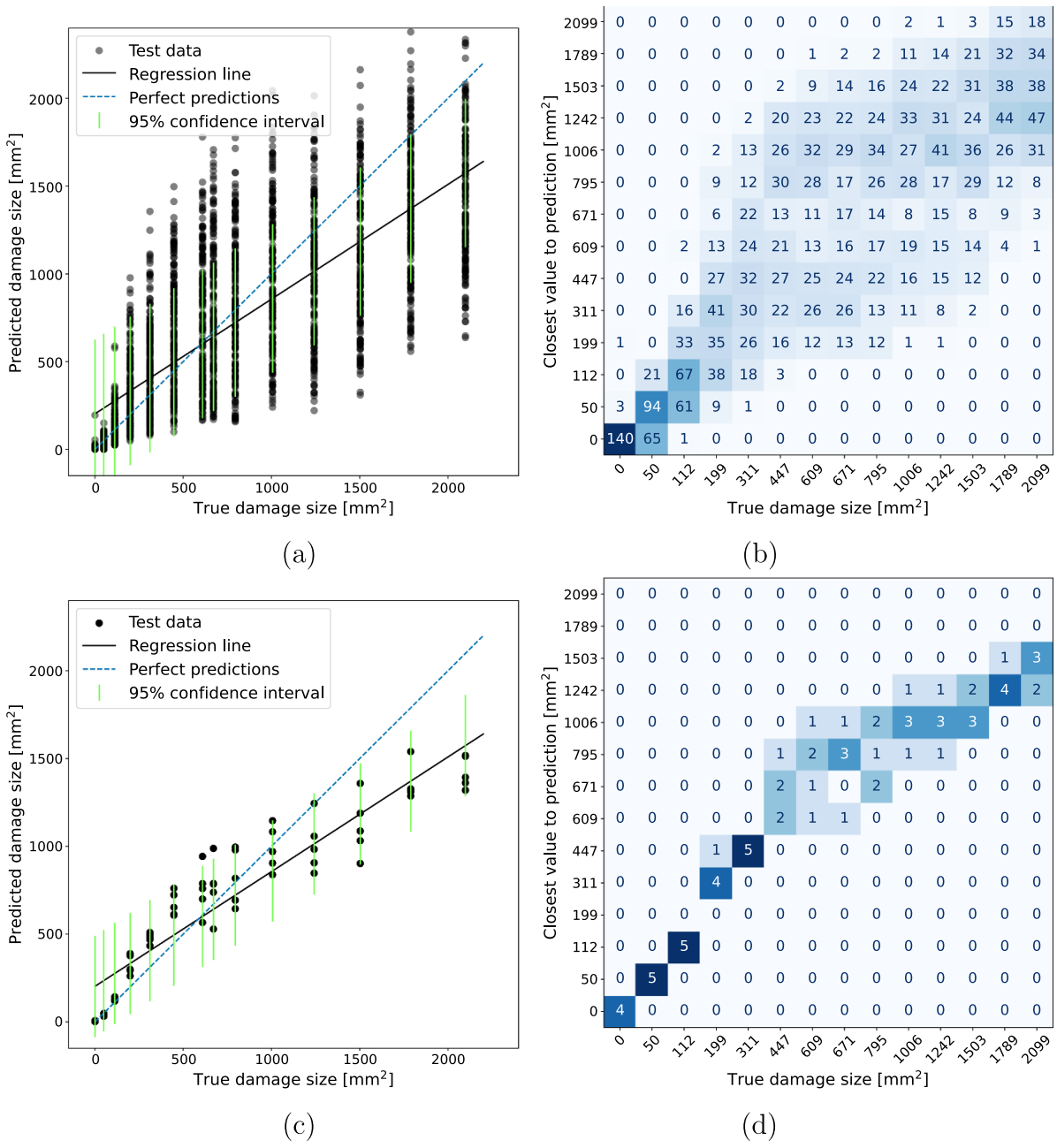


Figure 9. Predicted damage size with standard procedure (a), (b) and with ensemble voting (c), (d). The reduction of the variation in the predictions can be clearly seen in the scatter plots (left) and in the confusion matrices (right). The latter are generated by assigning all predictions made in a continuous spectrum to the nearest discrete class label.

detecting damage properly. The best options strongly depend upon the type of classification carried out. Most importantly, the performance for different damage locations changes and the variability depends upon the damage position. Despite this is an important influence factor when the training is performed in a generalized way (model to be tested with unknown measurements at unknown damage positions), the accuracy in damage detection and severity assessment is relatively high. In addition, results demonstrate that ensemble voting instead of single prediction can significantly improve performance in both binary and multi-class task.

While the binary classification accuracy estimating whether a damage is present or not is quite high, the severity assessment is generally more demanding and the accuracy evaluation requires a physics based interpretation to give meaning to an acceptable error in predicting damage size. To this end, various metrics are introduced to bound the confidence of the estimator. The results show once again that ensemble voting is essential to reduce the prediction variability and achieve a reasonable accuracy. In addition, the interpretation of results should be driven by physics informed reasoning defining a region of correct identification. Having and accounting an

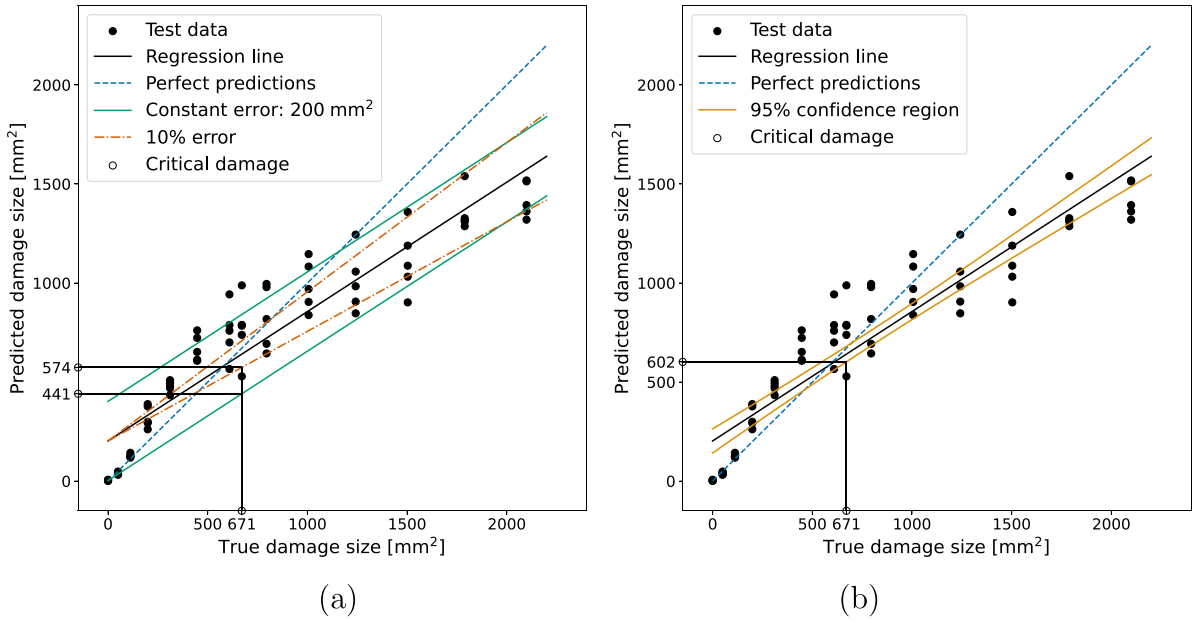


Figure 10. Calculation of the critical damage size using constant or relative error bounds (a) or with the confidence region (b).

acceptable error provides better accuracy and makes the predictor interpretable and explainable. Overall, this paper paves the way towards a unified approach for damage detection, size prediction and reliability assessment using CNN.

All this requires the definition of a reliability procedure in place of the mere accuracy analysis. However, a specific procedure is needed to fit the application to learning approaches. In addition to that, it is worth assessing the effect of other influence factors, such as temperature and loads, which are connected with the environment rather than inherent to the SHM system. Finally, the reduction of training resources can pave the way towards the implementation of such an approach on the edge, optimizing the data pipeline further. Future investigations are planned to encompass these key aspects enabling the deployment of AI based SHM.

Data availability statement

Data is already available here: http://openguidedwaves.de/. The data that support the findings of this study are available upon reasonable request from the authors. http://openguidedwaves.de/downloads/.

Acknowledgments

The authors gratefully acknowledge *DFG* and the members of the Scientific Network *Towards a holistic quality assessment for guided wave-based SHM* for the fruitful discussions (Project Number 424954879).

ORCID iDs

Oliver Schackmann https://orcid.org/0009-0005-8259-6681
Vittorio Memmolo https://orcid.org/0000-0003-1249-918X
Jochen Moll https://orcid.org/0000-0003-2299-2250

References

[1] Ciliberti D, Della Vecchia P, Memmolo V, Nicolosi F, Wortmann G and Ricci F 2022 The enabling technologies for a quasi-zero emissions commuter aircraft *Aerospace* **9** 319
[2] Mitra M and Gopalakrishnan S 2016 Guided wave based structural health monitoring: a review *Smart Mater. Struct.* **25** 053001
[3] Diogo A R, Moreira B, Gouveia C A J and Tavares J M R S 2022 A review of signal processing techniques for ultrasonic guided wave testing *Metals* **12** 936
[4] Cusati V, Corcione S and Memmolo V 2022 Potential benefit of structural health monitoring system on civil jet aircraft *Sensors* **22** 7316
[5] Lomazzi L, Giglio M and Cadini F 2023 Towards a deep learning-based unified approach for structural damage detection localisation and quantification *Eng. Appl. Artif. Intell.* **121** 106003
[6] Sattarifar A and Nestorovic T 2022 Emergence of machine learning techniques in ultrasonic guided wave-based structural health monitoring: a narrative review *Int. J. Progn. Health Manage.* **13** 1–29
[7] Lee J, Kim G, Ryu S and Park J 2022 Deep neural network-based structural health monitoring technique for real-time crack detection and localization using strain gauge sensors *Sci. Rep.* **12** 20204
[8] Shibu M, Kumar K P, Pillai V J, Murthy H and Chandra S 2023 Structural health monitoring using AI and ML based multimodal sensors data *Meas. Sens.* **27** 100762
[9] Malekloo A, Ozer E, AlHamaydeh M and Girolami M 2022 Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights *Struct. Health Monit.* **21** 147592172110368

[10] Mueller I *et al* 2023 Performance assessment for artificial intelligence-based data analysis in ultrasonic guided wave-based inspection: a comparison to classic path-based probability of detection *European Workshop on Structural Health Monitoring* and A Milazzo (Springer) pp 953–61
[11] Miorelli R, Fisher C, Kulakovskyi A, Chapuis B, Mesnil O and D'Almeida O 2021 Defect sizing in guided wave imaging structural health monitoring using convolutional neural networks *NDT&E Int.* **122** 102480
[12] Zhang S, Li C M and Ye W 2021 Damage localization in plate-like structures using time-varying feature and one-dimensional convolutional neural network *Mech. Syst. Signal Process.* **147** 107107
[13] U.S. Department of Defence USA 2009 MIL-HDBK No: 1823A. Non destructive evaluation system reliability assessment
[14] Géron A 2019 *Hands-On Machine Learning With Scikit-Learn, Keras and Tensorflow* (O'Reilly Media)
[15] Moll J, Kexel C, Kathol J, Fritzen C P, Moix-Bonet M, Willberg C, Rennoch M, Koerdt M and Herrmann A 2020 Guided waves for damage detection in complex composite structures: the influence of omega stringer and different reference damage size *Appl. Sci.* **10** 3068
[16] De Marchi L, Perelli A and Marzani A 2013 A signal processing approach to exploit chirp excitation in lamb wave defect detection and localization procedures *Mech. Syst. Signal Process.* **39** 20–31
[17] Michaels J E, Lee S J, Croxford A J and Wilcox P D 2013 Chirp excitation of ultrasonic guided waves *Ultrasonics* **53** 265–70
[18] Neuschwander K, Moll J, Memmolo V, Schmidt M and Bücker M 2019 Simultaneous load and structural monitoring of a carbon fiber rudder stock: experimental results from a quasi-static tensile test *J. Intell. Mater. Syst. Struct.* **30** 272–82
[19] Moll J, Kathol J, Fritzen C P, Moix-Bonet M, Rennoch M, Koerdt M, Herrmann A S, Sause M G and Bach M 2019 *Struct. Health Monit.* **18** 1903–14
[20] Bach M, Pouilly A, Eckstein B and MoixBonet M 2017 Reference damages for verification of probability of detection with guided waves *(Internationa Workshop on Structural Health Monitoring 2017)*
[21] McFee B, Raffel C, Liang D, Ellis D P, McVicar M, Battenberg E and Nieto O 2015 Librosa: audio and music signal analysis in python *Proc. 14th Python in Science Conf.* pp 18–25
[22] Akiba T, Sano S, Yanase T, Ohta T and Koyama M 2019 Optuna: a next-generation hyperparameter optimization framework *Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining)*
[23] Mueller I, Memmolo V, Tschöke K, Moix-Bonet M, Möllenhoff K, Golub M, Venkat R S, Lugovtsova Y, Eremin A and Moll J 2022 Performance assessment for a guided wave-based SHM system applied to a stiffened composite structure *Sensors* **22** 7529
[24] Tan M and Le Q V 2021 EfficientNetV2: smaller models and faster training (arXiv:2104.00298)