FIA 2020/22
XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC
Florianópolis, SC, Brasil

Machine Learning Techniques Comparison Applied to Structural Health Monitoring in Pipes

Oliveira, M.[1]**; Pazini, M. V. L.**[2]**;Corrêa, L. A.**[3]**; Clarke, T. G. R.**[4]

[1,2,3,4] Post-Graduation Program in Mining, Metallurgy and Materials Engineering (PPGEM), Federal University of Rio Grande do Sul (UFRGS), Porto Alegre - Brazil, {oliwer.marlon, marcuspazini, labcorrea, thomas.grclarke}@gmail.com

Abstract

Structural Health Monitoring systems using the ultrasonic guided wave method are able to identify, locate and estimate the severity of defects resulting from corrosive processes in structures. Such systems, when constituted of sensors permanently coupled or inserted in the structure and electronics of acquisition and transmission of data, are close to the concept of intelligent structures, which are those that have the ability to self-diagnose. One of the biggest problems for this approach is the sheer volume of data generated. The present work aims the elaboration of a methodology based on machine learning algorithms for processing data obtained from such systems. A synthetic database was produced for feed several algorithms such as kNN, CNN, MLP and SVM. A significantly high success rate was observed for the identification of the damaged structure, i.e., the method presented in this work proved to be very promising and at the moment a prototype is installed in a refinery, being tested under real operating conditions.

Keywords: Machine Learning, SHM, Guided Wave inspection, Pipeline inspection.

PACS: 43.20.Mv, 43.60.Pt, 43.60.Np.

1. INTRODUCTION

Machine Learning (ML) is the part of Artificial Intelligence that focus on pattern recognition [1]. It is divided in unsupervised and supervised approaches. Unsupervised algorithm are applied on data where the labels are unknown in order to find any pattern. Supervised algorithms are responsible for classifying data where the labels are well known.

The learning of supervised approaches can be model-based, example-based or through feature extraction. In this work we will use an example based method. Generally these methods depend on a large set of examples for which the parameter values are known. They deduce the label values for the input example from the known values in similar examples. Thus, it is assumed to have sufficient data to make such inference [2]. In [3] a similar approach is used for Structural Health Monitoring (SHM) systems. It was called population-based SHM and the main idea is to transfer, in some sense, information between groups of similar structures.

Machine Learning algorithms have been proposed in recent research as an alternative to statistical methods. Such algorithms have been successfully used in computer vision [4, 5]. However, its application is not limited to images and videos. Recently engineers are considering these techniques in SHM applications as well [6].

SHM systems based on Guided Waves (GW) are becoming research focus because they can significantly reduce the inspection costs of safety-critical structures, e.g. aerospace, nuclear, oil and gas industries. SHM systems can be consider effective if it combines good sensitivity to defects and low cost [7–10].

The identification problem in SHM systems can be seen as a hierarchical structure, such as the following levels [11]:

- Level 1: detection. Detect if there is defect in the structure.
- Level 2: localisation. Informs about the probable position of the damage.
- Level 3: classification. Informs about the type of the damage.
- Level 4: assessment. Estimates the extent of the damage.
- Level 5: prediction. Informs about the safety of the structure.

Guided Waves for SHM systems were used in [12]. Features were extracted with and without baseline, totalling 167 features. Three classification algorithms for detecting damage were applied: adaptive boosting (AdaBoost), SVM, and a method combining the two (AdaSVM). Accuracies were 87.7%, 92.5% and 93.5%, respectively. The application was in natural gas pipelines.

Deep neural network was applied over GW for SHM. Tests were performed over four different plates, called A, B, C and D. Different classifiers were trained, e.g. training with 4, 3, 2 and only 1 of them. The results showed that deep learning could achieve a very high accuracy, 99%+ with training data of only one plate, against SVM which achieved only 62% [13]. However, this idea is criticised in [14, 15], where the previous mentioned idea could be over-engineering a problem that is well-defined on less input data. It is a common critique of ML by classical statistics and meticulous data science.

Guided Waves were applied in [16] to SHM in pipelines. Principal Component Analysis (PCA), hierarchical clustering (HC) and multinomial logistic regression (MLR) were used for data-driven semi-supervised and supervised learning algorithms for health monitoring of these pipes. The semi-supervised approach detects if the structure is damaged or not by using a hierarchical clustering based algorithm. The supervised learning approach hence the localization in a multinomial logistic regression framework. Both experiments were validated and demonstrated that the fully data-driven techniques could precisely detect and localize cracks on two cast iron pipes of different lengths.

In this paper we focus on ML supervised techniques, once the defects were created in a synthetic form and we know them. We will use

SHM levels 1, 2 and parts of 3 and 4. It is explained in Section 3 with the ML algorithms and their parameters. Section 2 describe our hybrid dataset used in this paper. Section 4 presents the results and comments. Finally, Section 5 shows the conclusions.

2. DATASET

Datasets play a very important role in any Machine Learning application. In some specific situations the acquisition of the data could be expensive in time and/or resources, e.g. experiments with guided waves. In order to overcome this problem it is common the use of numeric models to produce synthetic data. However, some environmental changes, such as noise and temperature are difficult to reproduce by numerical model. Finally, hybrid approaches, being a combination of real data with synthetic data, appears to be a good manner to overcome the difficulty of gather real data and the weakness of the synthetic data.

In this work we use a hybrid dataset, combining a finite element model of a damaged structure pipe and a multiple environmental cycles of an undamaged structure. The experimental and numerical setup are explained in the Section 2.1 with the synthetic experimental-numerical database (more details and an evaluation of these methods can be found in [17]).

2.1 Experimental Setup

The experimental setup consists in a undamaged 8" steel pipe schedule 40 with 6.48 meters of length as sample. In order to perform the guided wave test we attached piezoelectric transducers to build the guided wave collar. It was composed by 32 shear plates (3x13x1 mm^3 manufactured by PI [18] with PIC255 as material). They were divided in two rings (16 shear plates in each) spaced by 20 mm and equally divided in the circumference. They were brazed in a flexible PCB and then glued in the pipe. A homemade multiplexer and amplifier were used to perform pulse-echo and pitch-catch configuration (details about this hardware used can be found in [19]). This configuration allows us to perform an inspection using the T(0,1) guided wave mode [19–23]. Figure 1 shows a schema of the experiment, where the grey triangles show the longitudinal position of the defects, at the edge of the pipe the GW collar.

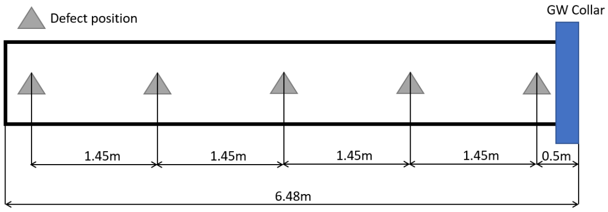

Figure 1: Experimental scheme of the experiment. In blue the position of the GW collar. The grey triangles show the position of the syntactical defects in the sample.

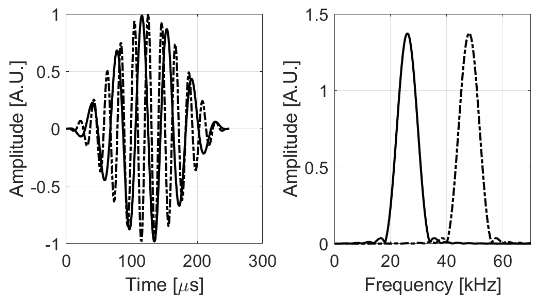

Figure 2: Wave-forms used in the wave guide excitation. The solid line represents the waveform (i): 26 kHz with 6.5 cycles. The dashed line represents the wave-form (ii): 48 kHz with 12 cycles. The left figure presents the time domain representation and the right figure the frequency domain representation.

The waveform used was a tone-burst signal. It is described by the follow equation:

$$A(t) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi t}{N}\right)\right]\sin(2\pi f_c t), \quad (1)$$

where t is time in seconds, N is the number of periods of the wave, and f_c is the central frequency of the wave. Two different configurations were used in this experiment: (i) 26 kHz with 6.5 cycles; (ii) 48 kHz with 12 cycles. Figure 2 presents the waveform used for the excitation of the sensors.

The dataset used in the present work consists only of the data that come from the positive axis of the pipe (left side in the Figure 1). The setup configuration allows us to select the direction of the recorded signal. This is a standard procedure in pipeline inspection [24]. The test were made during one day to gather variability in the collected signals (electrical noise and modal noise). The temperature variation was between 23.29^{o}C and 26.83^{o}C.

2.2 Numerical setup

The numerical setup was built to recreate the experimental setup described in Section 2.1. The dynamic equilibrium equation was solved using finite element method. An explicit schema for the integration of the equilibrium equation was used. The CFL was set to 0.2 to ensure the stability of the procedure. One hexahedral mesh was used in the discretisation. It is composed by 540 linear elements in the length, 4 linear elements in the radius, and 64 linear elements in the circumference. The input force assign to 16 nodal forces equally spaced in a tangential direction to mimic the experimental test. The amplitudes of those forces were presented in Figure 2. The material was with an isotropic one with: $E = 217$ GPa, $\nu = 0.2896$, and $\rho = 7923\text{kg/m}^3$.

Regarding the defect, we introduced in the numerical model a total of 3,000 different damages. Each defect was parameterized by four parameters. They are the following: (i) the severity of the depth (0.0625% up to 0.75% of the thickness, 12 steps); (ii) the angle extension (3deg up to 30deg, 5 steps); (iii) longitudinal defect extension (0.012 m up to 0.24 m, 5 steps); (iv) longitudinal position (0.5 m up to 6.3 m, 5 steps). The shape (half period 2D sinus) of this defect can be seen in Figure 3. These damages gave origin to the case scenarios showed in Section 3 Table 1.

2.2.1 Creation of the dataset

In SHM often a large amount of collected data is processed. Most part of the collected data come from situations where there is no damage and the structure is considered to operate normally. ML classifiers are generally supervised learning, requiring examples of both classes. Thus, the lack of damaged data from a structure might present a challenge to the task of training a good classifier.

In order to generate an adequate dataset to be utilised to train a ML classifier, we merged different base signals (signals with no defect) with the defects created via numerical simulations.

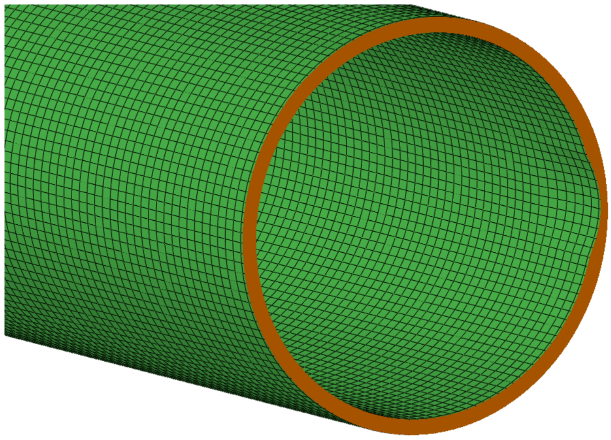

Figure 3: Clip of the mesh used in the numerical models. The cut plane presented shows the reduction of the cross section for one case.

Therefore, artificially increasing the ratio between damaged and undamaged signals present in the database.

Prior merging the numerical and the experimental signals a number of operations are needed to ensure that the numerical data and the experimental sample are correctly added, as the environmental conditions often affect the frequency and amplitude of the experimental data. Thus, the numerically generated data has to be modified to match the experimental, accordingly. In order to match the signals we needed three steps: first, the amplitude from the model is corrected to match the amplitude from the edge reflection of the experimental base; second, the crosstalk from the selected sample is cut from the signal, and third step is to perform upper sampling and temperature compensation on the numerical signal to adjust the signal length to the base signal.

Once the previous steps are performed the experimentally obtained data and the numerical induced defect are added together in order to generate a new signal. This signal consists of a base signal with a numerically introduced defect. A simplified scheme of the database creation is shown in Figure 4, where the processing steps refer to the first, the second and the third previously mentioned operations.

When those steps are applied to every possible combination of the n numerically generated data and the m experimentally collected data. Thus, is possible to generate a total of $n*m$ synthetic samples that are a unique combinations between nu-

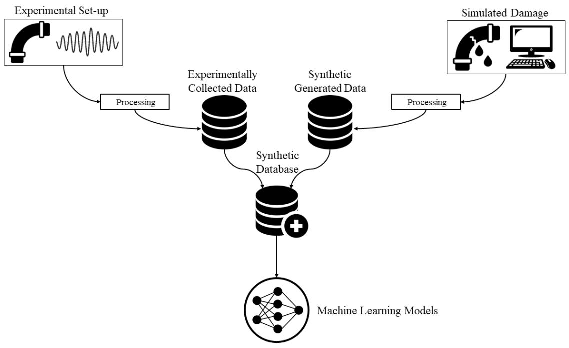

Figure 4: Hybrid dataset creation flowchart.

merical model data and experimental data. Therefore, the ratio between samples with defects and base signals always follows the Equation 2.

$$N = \frac{n(m-p)}{n(m-p)+m}, \quad (2)$$

where p is the number of bases where numerical defects were not added yet.

In our case, the baseline (samples without defect) is composed by 90 samples. We chose 90 numerical models randomly with the same defect position and we repeated this for 5 different positions (as showed in Figure 1). The complete dataset has 540 guided wave test. However, it is still a small sample.

Based on the work made by Ewald, Groves, and Benedictus [25] we added white noise in order to increment the size of our dataset and increase the variability [26]. Three different levels were used: 2.5dB, 4.4dB and 6dB. Those levels were choose arbitrarily. At this point we should stress that the presence of white noise can lead to the inverse crime [27–29] and the use of white noise is not a consensus in the domain [30] (even with a solid used in system identification [31, 32]).

2.2.2 Feature extraction

Guided wave pipe modes have a distribution on the circumference and this can be measured through a sensor ring. These distributions follow the restriction of periodicity imposed by the pipe. Therefore, the wavelengths in circumferential direction are always whole fractions of the perimeter. Thus, the F(1, m) modes, where mode order is 1, will have only one wavelength arranged in the perimeter, the modes F(2, m) will have two, and so on [20]. In order to find the circumference order of guided wave modes we can postulate modes are real functions. In [17] the same decomposition was used as features and lately used to feed a PCA and Independent Component Analysis (ICA) to retrieve the damage.

An arbitrary real function could be decomposed into an infinite sum of cosines through the Fourier Transform. For discrete signals (when only function values exist for discrete instants), the Fourier Transform should be made using a discrete implementation, the discrete Fourier transform (DTF). The DTF presents a particular case when the discrete function interval input is considered periodic.

The result of the DFT is a sequence of values of the same input string size, representing the coefficients whose periods are multiple integers of the signal length [33, 34]. If we change the implicit idea of applying Fourier in time to decompose frequencies and pass for the spatial application we can decompose the signal into the number of wavelengths within the perimeter what will be representing the circumferential distribution of the modes. This application is given in the direction of n in the data matrix M_{nm}, where n is the number of sensors and m is the number of time samples. Thus, the decomposition by DFT reveals the quantities for each circumferential order at each instant of time. In this work we used the Fastest Fourier Transform in the West (FFTW) implementation [35]. DFT is defined by the transform and inverse pair:

$$DFT\{x_n\} = X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi\frac{kn}{N}}, \quad (3)$$

$$DFT\{x_n\}^{-1} = x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{i2\pi\frac{kn}{N}}, \quad (4)$$

where X_k is the DFT of the real signal x_n, N is the number of samples in the direction n of the data matrix, k is the circumferential wave-number, and i is the complex number. If we take a vector a_n composed of one of the m columns of M, that is, a sample time for all channels and applying DFT we will have:

$$A_k = DFT\{a_n\}, \quad (5)$$

The elements of A_k will be the coefficients that will represent up to the circumferential order $N/2$. In order to get only real components the Discrete Hilbert Transform (DHT) is calculated through its implementation in the frequency domain where the core of the Hilbert transform is H_k [33]. Thus:

$$DHT\{x_n\} = \hat{x}_n = DFT^{-1}\{H_k DFT\{x_n\}\}, \quad (6)$$

where $\hat{x}_n$ is the DHT of x_n. The kernel is defined as follow:

$$H_k = \begin{cases} 0, & \text{if} \quad k = 0 \\ i, & \text{if} \quad 0 < k < N/2 \\ 0, & \text{if} \quad k = N/2 \\ -i, & \text{if} \quad N/2 < k < N. \end{cases} \quad (7)$$

The process is repeated for the m columns of M obtaining $\hat{M}_{km}$ presenting the decomposed signals of the time domain modes [36].

3. METHODOLOGICAL PROCEDURES

In this work we used four different machine learning algorithms for damage classifications. They are: k-Nearest Neighbour (k-NN), Support vector machines (SVM), Multilayer Perceptron (MLP) and Linear Discriminant Analisys (LDA). These were chosen because they are popular in the field of pattern recognition in recent years. They are not very complex and not computationally expensive.

k-NN is an algorithm widely used for pattern recognition. The input is the k closest training examples in the feature space. The output is the class that an item is classified by the most voted from neighbours, k is a low positive integer [37].

The most common used distance metric in k-NN is Euclidean.

$$d(a,b) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2}.$$

Where a and b are two points in the space, $x_1, x_2, x_3 \ldots x_n$ are the coordinates of a and $y_1, y_2, y_3 \ldots y_n$ are the coordinated of b.

Apart from Euclidean distance we tried manhattan, chebyshev and minkowski distances as well. In addition, we tried different number of neighbours.

SVM are one classification method that uses ML to increase the accuracy of prediction. In SVM the training data may be linearly separable in the input spaces or not linearly separable, in this last case they map the input space into a high-dimensional feature space to obtain linear separation.

The hyperplane separates the training data by a maximal margin. Vectors lying on one side of the hyperplane are classified as -1 and vectors lying on the other side are classified as 1. The training instances closest to the hyperplane are the *support vectors* [38].

In this paper we will try different kernels for SVM, such as *linear*, *radial basis function*, *poly* and *sigmoid*.

MLP is one type of feedforward artificial neural network. It consists of layers of nodes (neurons) that uses a non-linear activation function. The output of a node is scaled by the connecting weight and fed forward to be an input to the next layer neuron. MLP is able to classify data that is not linear, because its multiple layers and non-linear activation [39].

Activation functions are not limited to: *identity* useful to implement linear bottleneck $f(x) = x$; *logistic* sigmoid $f(x) = 1/(1+exp(-x))$; *tanh* hyperbolic tan function $f(x) = tanh(x)$ and *relu* the rectified linear unit function $f(x) = max(0,x)$ [40].

MLP utilises a technique called backpropagation for training. Where the weights in the network are initially set to low random values and thus the algorithm computes the local gradient of the error surface and update the weights in the way of steepest local gradient [40].

It is known that this approach leads to a *false* result. The use of noisy data can mask the *real* algorithms performance once occurs the inverse crime. The phase information in the signal cannot be reconstructed using MLP - normally neural networks are able to reconstruct phase information. In [41] is presented these neural network problems together other related problems such as input data correlation. The phase contains several information needed in guided waves for non destructive testing.

LDA creates an optimal linear discriminant function $f(x) = W^T x$ mapping the input into the classification space. The class identification of this sample is based on some sort of metric, e.g. Euclidean distance. A classic LDA training stage is created with within and between-class scatter matrices analysis [42]. Within and between-class scatter matrices are calculated as follows:

$$S_w = \frac{1}{M}\sum_{i=1}^{M} Pr(C_i)\Sigma_i, \tag{8}$$

$$S_b = \frac{1}{M}\sum_{i=1}^{M} Pr(C_i)(m_i - m)(m_i - m)^T, \tag{9}$$

where S_w is the within-class scatter matrices and S_b is the between-class scatter matrices. $Pr(C_i)$ is the prior probability of class C_i; m is the overall mean vector; Σ_i is the average scatter of the sample vectors of distinct classes C_i around the representative mean vector m_1.

LDA is different from PCA, because the maximum dimensionality that can be achieved is the number of classes minus 1. Whereas in PCA is the number of classes [43].

In this work we will explore solvers *lsqr* and *svd* only.

3.1 Machine Learning Application

In order to apply machine learning algorithms over our data we created 4 different scenario cases. These are the following: case 1 is a binary classification, just for classify if the sign has or has not defect. Case 2 classifies the position of the defect. Case 3 classifies the severity of the damage, the loss of area of the cross section of the pipe, what we call Cross Section Area (CSA).

Table 1: Scenario cases for classification

Case 1	Binary classification
Case 2	Position of the defect (5 classes)
Case 3	CSA
Case 4	Position and CSA

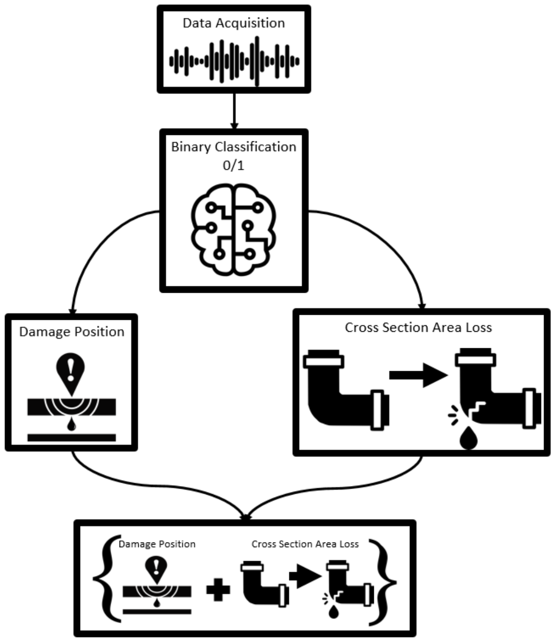

Figure 5: Workflow for the feature extraction

Finally, case 5 classifies the position and the CSA together. Table 1 shows the scenario cases. These scenarios were obtained by the damages presented in the Section 2.2.

The input, feature extraction and the classifications according to the case scenarios are show in Figure 5.

In Case 1 we used 450 ($90 \times 5 positions$) data with defect and 90 data without any defect, resulting in 540 data. It before applying AScan and before adding noise.

We used different levels of noise, such as 0, 0.33, 0.66 and 1 Db, increasing our dataset four times. Extracting AScan features gives us 9 different channels for each signals, as we increased it 4 times with noise, in total we have $540 \times 9 \times 4 = 19,440$. Figure 6 shows one signal (from one AScan channel) before and after applying different levels of noise. The solid line represents the absence of noise, the dash-dot line (-.)contains white noise of 0.33dB, the dotted line (..) noise of 0.66dB and the dashed line (–) line noise of 1dB.

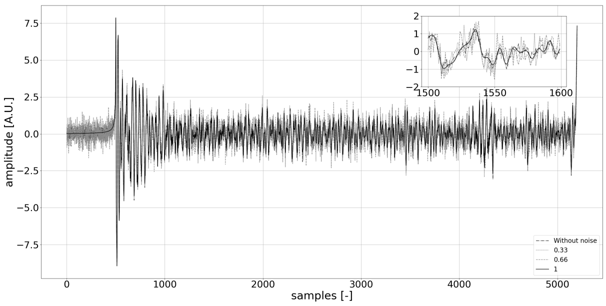

Figure 6: Original signal with different noise levels added.

Case 2, 3 and 4 ignores all the data without defect. It means 3,240 less data or a total of 16,200 data.

Case 2 contains damage in 5 different positions. It is the position of the damage from the sensor in metres, it gives us 5 classes being labeled ad 0.5, 1.95, 3.4, 4.85 and 6.3. Each damage position appears 90 times before adding noise. The histogram of the classes distribution can be visualised in Figure 7. Note that each position plus the non defective has the same number of occurrence in the dataset.

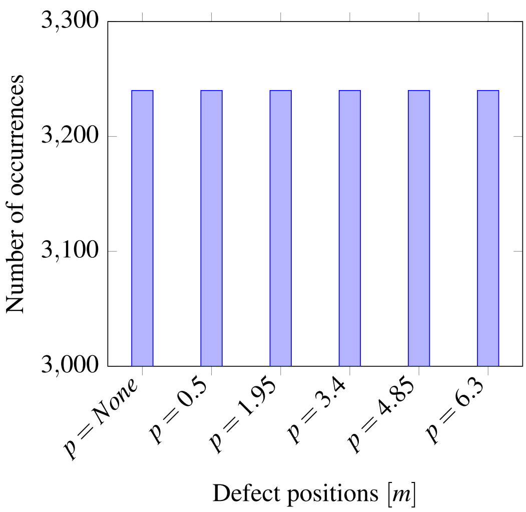

Figure 7: Histogram of the damage positions

CSA or damage severity (Case 3) is the most diverse class of our data, it means a high number of possibilities (222 to be more precise). In order to group all these classes we created the following criteria: smaller than 2.20% in one class (csa = 1), between 2.20% and 6.00% in other class (csa = 2) and finally damage greater than 6.00% in the third class (csa = 3). The histogram of the distribution of the CSA in 3 levels is showed in Figure 8.

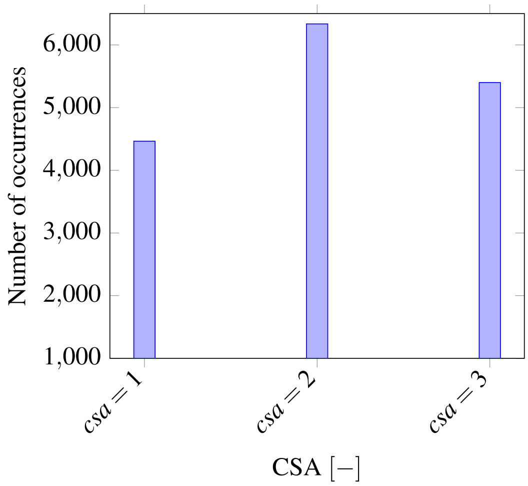

Figure 8: Histogram of labels from the Case 3

Finally we tried to classify position and sizing together. In total we have 15 different labels, being 5 classes for position and 3 classes for CSA. It is our Case 4.

Before applying any ML algorithm over our data we had to split or data in training and testing. It because we are using a supervised approach. Thus, we split or dataset in 70% for training and 30% for testing, randomly and with no cross validation. In total we have 13,608 data for training and 5,832 data for testing.

4. RESULTS

In this experiment we used the following parameters: k-NN with Euclidean distance and 1 neighbour. MLP used activation function *relu* and solver *adam*, with 1 hidden layer, 50 hidden units and number maximum of iterations equal to 100. LDA used solver svd - singular value decomposition. SVM is using kernel *linear*.

4.1 Damage identification

The first stage of our SHM proposed system is to identify if a structure has or has not damage. In order to classify this, we used class containing 0 for non damaged and 1 for damaged. Binary classifier was used. It was called Case 1 and the results are shown in Table 9. Parameters have not been tuned for this experiment, because it is a trivial problem and we obtained very high accuracy.

Table 2: kNN with different *k* values

k	Accuracy
1	**0.6359**
3	0.4194
5	0.3981
7	0.3768

Table 3: kNN with k = 1 and different distance metrics

Metric	Accuracy
minkowski	0.6359
euclidean	**0.6359**
manhattan	0.6182
chebyshev	0.4050

4.2 Damage localization

The following experiments is applied to Case 2 and try to classify the position of and identified damage. For this experiment only structures classified as damaged were considered.

As we have 5 different damage positions it is not a trivial problem any more. Thus, we need tuning the parameter for the ML algorithms. In kNN we tried different number of *k*, see Tables 2 and 3. LDA we tried different solvers, see Table 4. MLP we tried different activation functions, solvers, number of hidden layers and number of hidden units (neurons), see Tables 5, 6 and 7. SVM we tried different kernels, see Table 8.

Finally, after tuning the parameters we can approximate the optimal accuracy for damage identification, localization, sizing and localization + sizing. Results are shown in Table 9

4.3 Damage localization + sizing

Table 9 shows the results for all the four cases. Results shown are based in the tuning made in the Position Classification (Case 2). Thus, we used k-NN with only one neighbour and MLP with one layer and 50 units, SVM with polynomial kernel.

It is clear that k-NN has the best accuracy com-

Table 4: LDA with different solvers

Solver	Accuracy
svd	0.2647
lsqr	0.2647

Table 5: MLP with different activation functions and solvers, with one hidden layer and 50 neurons

Activation	Solver	Accuracy
relu	adam	0.3530
relu	sgd	0.3655
relu	lbfgs	**0.4010**
tanh	lbfgs	0.3690
tanh	sgd	0.3959
tanh	adam	0.2864
identity	lbfgs	0.2901
identity	sgd	0.2768
identity	adam	0.2690
logistic	lbfgs	0.3501
logistic	sgd	0.3895
logistic	adam	0.3329

Table 6: MLP with different number of hidden units (neurons) for activation *relu* and solver *adam*

Units	Accuracy
25	0.3407
50	0.3530
75	0.3532
100	0.3516
150	**0.3542**
200	0.3528

Table 7: MLP with different number of hidden layers and units (neurons) for activation *relu* and solver *adam*

Layers	Units	Accuracy
1	50	0.3530
2	50, 50	0.4168
3	50, 50, 50	0.4133
4	50, 50, 50, 50	0.4141
4	50, 100, 150, 200	0.4178
2	100, 50	0.4248
3	150, 100, 50	0.4285
4	200, 150, 100, 50	0.4385
4	200, 150, 100, 50	0.4366
5	250, 200, 150, 100, 50	**0.4393**

Table 8: SVM with different kernels

Kernel	Accuracy
linear	0.2841
poly	**0.3516**
rbf	0.3350
sigmoid	0.2064

Table 9: Accuracy for different algorithms

Algorithm	Case 1	Case 2	Case 3	Case 4
kNN	0.1000	0.6359	0.7159	0.5795
LDA	0.9949	0.2841	0.3834	0.2665
MLP	0.1000	0.4393	0.5312	0.2210
SVM	0.9993	0.1146	0.4934	0.2476

pared to the other algorithms in all the four cases. Being nearly 100% for case 1, around 63% for case 2, 69% for case 3 and finally 57% for case 4. Case 4 is the most difficult to classify because combine the classes of positions (5) and classes of CSA (3), totalling 15 classes.

It is of common knowing that neural networks discard some information of the signal, depending on the activation function used and the normalization of the data. We tried different combinations in this work. However, k-NN still overcame accuracy in all the cases.

Finally, according to [44] the 1-Nearest Neighbour classifier is the most complex nearest neighbour model. It can lead to over-fit the model. It may explain why we obtained a higher accuracy with 1-Nearest Neighbour and we can conclude that it may not be the classifier that best represents our model. Another explanation could be that the added noise helped the 1-Nearest Neighbour more than the other algorithms.

5. CONCLUSION

In this work we provided the application of machine learning algorithms over a hybrid dataset. The dataset used consists of signal from guided waves signal obtained by sensors attached on a steel pipe, combining both synthetic and experimental.

Features were extracted followed by the application of supervised machine learning algorithms in order to classify different level of defects, such as defective/non defective, the position estimates and the severity of the damage.

We tested four different machine learning algorithms, k-nearest neighbour (kNN), multilayer perceptron (MLP), support vectors machine (SVM) and linear discriminant analysis (LDA). Results have shown accuracy of 100% for binary classification in most algorithms. Position estimation and severity of the damage k-NN have shown the best accuracy using kNN algorithm being 63% and 71% respectively.

In the future we intend to use an automated algorithm to tune the hyper-parameters of the machine learning algorithms, specially for MLP, and, applying deep learning algorithms as convolutional neural network is forecast.

REFERENCES

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. *Proc. Ninth IEEE Int. Conf. Comput. Vis.*, (April):0–11, 2003. doi:10.1109/ICCV.2003.1238424.

[3] LA Bull, PA Gardner, J Gosliga, TJ Rogers, N Dervilis, EJ Cross, E Papatheou, AE Maguire, C Campos, and K Worden. Foundations of population-based shm, part i: Homogeneous populations and forms. *Mechanical Systems and Signal Processing*, 148:107141, 2020.

[4] Marlon Oliveira, Houssem Chatbri, Ylva Ferstl, Suzanne Little, Noel E. O'Connor, and Alistair Sutherland. Irish sign language recognition using principal component analysis and convolutional neural networks. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, Nov 2017. doi:10.1109/DICTA.2016.7796984.

[5] Marlon Oliveira, Houssem Chatbri, Suzanne Little, Noel E. O'Connor, and Alistair Sutherland. A comparison between end-to-end approaches and feature extraction based approaches for sign language recognition. In *International Conference on Image and Vision Computing New Zealand (IVCNZ), 2017*, pages 1–6. IEEE, 2017.

[6] Majdi Flah, Itzel Nunez, Wassim Ben Chaabene, and Moncef L Nehdi. Machine learning algorithms in civil structural health monitoring: a systematic review. *Archives of Computational Methods in Engineering*, pages 1–23, 2020.

[7] Thomas Clarke, Peter Cawley, Paul David Wilcox, and Anthony John Croxford. Evaluation of the damage detection capability of a sparse-array guided-wave SHM system applied to a complex structure under varying thermal conditions. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 56(12):2666–2678, 2009. ISSN 08853010. doi:10.1109/TUFFC.2009.1357.

[8] Ajay Raghavan. *Guided-wave structural health monitoring*. PhD thesis, 2007.

[9] Mira Mitra and S Gopalakrishnan. Guided wave

based structural health monitoring: A review. *Smart Materials and Structures*, 25(5):053001, 2016.

[10] Christoph Kralovec and Martin Schagerl. Review of structural health monitoring methods regarding a multi-sensor approach for damage assessment of metal and composite structures. *Sensors*, 20(3):826, 2020.

[11] Charles R Farrar and Keith Worden. *Structural health monitoring: a machine learning perspective*. John Wiley & Sons, 2012.

[12] Yujie Ying, Joel Harley, James H. Garrett, Yuanwei Jin Jr., Irving J. Oppenheim, Jun Shi, and Lucio Soibelman. Applications of Machine Learning in Pipeline Monitoring. *International Workshop on Computing in Civil Engineering*, (2011):194–201, 2011.

[13] Joseph Melville, K. Supreet Alguri, Chris Deemer, and Joel B. Harley. Structural damage detection using deep learning of ultrasonic guided waves. *AIP Conference Proceedings*, 1949(April), 2018. ISSN 15517616. doi:10.1063/1.5031651.

[14] Arnaud Mignan and Marco Broccardo. One neuron versus deep learning in aftershock prediction. *Nature*, 574(7776):E1–E3, 2019.

[15] Arnaud Mignan and Marco Broccardo. A deeper look into 'deep learning of aftershock patterns following large earthquakes': Illustrating first principles in neural network physical interpretability. In *International Work-Conference on Artificial Neural Networks*, pages 3–14. Springer, 2019.

[16] Debarshi Sen, Amirali Aghazadeh, Ali Mousavi, Satish Nagarajaiah, Richard Baraniuk, and Anand Dabak. Data-driven semi-supervised and supervised learning algorithms for health monitoring of pipes. *Mechanical Systems and Signal Processing*, 131:524–537, 2019.

[17] Chang Liu, Jacob Dobson, and Peter Cawley. Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2199):20160736, 2017.

[18] Pi ceramic. https://www.piceramic.com/en/. Accessed: 2020-11-04.

[19] Ricardo C Jacques, Henrique H de Oliveira, Rafael WF dos Santos, and Thomas GR Clarke. Design and in situ validation of a guided wave system for corrosion monitoring in coated buried steel pipes. *Journal of Nondestructive Evaluation*, 38(3): 65, 2019.

[20] Mike JS Lowe, David N Alleyne, and Peter Cawley. Defect detection in pipes using guided waves. *Ultrasonics*, 36(1-5):147–154, 1998.

[21] MHS Siqueira, CEN Gatts, RR Da Silva, and JMA Rebello. The use of ultrasonic guided waves and wavelets analysis in pipe inspection. *Ultrasonics*, 41 (10):785–797, 2004.

[22] RC Jacques, HTH De Oliveira, RWF Dos Santos, and TGR Clarke. Design of an emat guided wave collar for coated riser inspection (march 2020). *IEEE Sensors Journal*, 2020.

[23] Yihui Da, Bin Wang, DZ Liu, and Zhenghua Qian. An analytical approach to reconstruction of axisymmetric defects in pipelines using t (0, 1) guided waves. *Applied Mathematics and Mechanics*, 41 (10):1479–1492, 2020.

[24] Paul D Wilcox. Guided-wave array methods. *Encyclopedia of structural health monitoring*, 2009.

[25] Vincentius Ewald, Roger M. Groves, and Rinze Benedictus. DeepSHM: a deep learning approach for structural health monitoring based on guided Lamb wave technique. page 19, 2019. ISSN 1996756X. doi:10.1117/12.2506794.

[26] Jesper Sören Dramsch. Machine learning in 4d seismic data analysis: Deep neural networks in geophysics. 2019.

[27] Armand Wirgin. The inverse crime. *arXiv preprint math-ph/0401050*, 2004.

[28] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: discretization, model reduction and inverse crimes. *Journal of computational and applied mathematics*, 198(2):493–504, 2007.

[29] Carlos Eduardo Chávez, Felipe Alonzo-Atienza, and Diego Alvarez. Avoiding the inverse crime in the inverse problem of electrocardiography: estimating the shape and location of cardiac ischemia. In *Computing in Cardiology 2013*, pages 687–690. IEEE, 2013.

[30] Mahjoub El Mountassir, Slah Yaacoubi, and Fethi Dahmene. Reducing false alarms in guided waves structural health monitoring of pipelines: Review synthesis and debate. *International Journal of Pressure Vessels and Piping*, page 104210, 2020.

[31] Panos Z Marmarelis and Vasilis Z Marmarelis. The white-noise method in system identification. In *Analysis of physiological systems*, pages 131–180. Springer, 1978.

[32] Lennart Ljung. System identification. *Wiley encyclopedia of electrical and electronics engineering*, pages 1–19, 1999.

[33] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[34] Simon Haykin and Barry Van Veen. *Signals and systems*. John Wiley & Sons, 2007.

[35] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[36] Henrique Tormen Haan de Oliveira. Projeto de um colar de ondas guiadas para aplicação em tubulação enterrada. Master's thesis, Universidade Federal do Rio Grande do Sul, 2017.

[37] R Naoum, HH Owaied, and S Joudeh. Development of a new Arabic sign language recognition using k-nearest neighbor algorithm. *J. Emerg. Trends Comput. Inf. Sci.*, 3(8):1173–1178, 2012. URL http://isite.persiangig.com/Files/KNN/Arabic{_}Naoum.pdf.

[38] Simon Tong and Daphne Koller. Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res.*, pages 45–66, 2001. ISSN 15324435. doi:10.1162/153244302760185243.

[39] Giorgio Licciardi, Ruggero Giuseppe Avezzano, Fabio Del Frate, Giovanni Schiavon, and Jocelyn Chanussot. A novel approach to polarimetric SAR data processing based on Nonlinear PCA. *Pattern Recognit.*, 47(5):1953–1967, 2014. ISSN 00313203. doi:10.1016/j.patcog.2013.11.009.

[40] M.W Gardner and S.R Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.*, 32(14-15):2627–2636, 1998. ISSN 13522310. doi:10.1016/S1352-2310(97)00447-0. URL http://www.sciencedirect.com/science/article/pii/S1352231097004470.

[41] Jesper Sören. General rights Machine Learning in 4D Seismic Data Analysis: Deep Neural Networks in Geophysics. 2020. URL https://www.fysik.dtu.dk.

[42] W Zhao, R. Chellappa, and S. Krishnaswamy. Discriminant analysis of principal components for face recognition. *Face Recognit. From Theory to Appl.*, pages 73–85, 1998. ISSN 1543-5474. doi:10.1109/AFGR.1998.670971. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=670971.

[43] Oya Aran and Lale Akarun. A multi-class classification strategy for Fisher scores: Application to signer independent sign language recognition. *Pattern Recognit.*, 43(5):1776–1788, 2010. ISSN 00313203. doi:10.1016/j.patcog.2009.12.002.

[44] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.