



Tu, L., Pyle, R., Croxford, A. J., & Wilcox, P. D. (2022). Potential and limitations of NARX for defect detection in guided wave signals. *STRUCTURAL HEALTH MONITORING*.  
<https://doi.org/10.1177/14759217221113240>

Publisher's PDF, also known as Version of record

License (if available):  
CC BY-NC

Link to published version (if available):  
[10.1177/14759217221113240](https://doi.org/10.1177/14759217221113240)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

This is the final published version of the article (version of record). It first appeared online via SAGE Publishing at <https://doi.org/10.1177/14759217221113240>. Please refer to any applicable terms of use of the publisher.

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

# Potential and limitations of NARX for defect detection in guided wave signals

Xin L Tu , Richard J Pyle , Anthony J Croxford   
and Paul D Wilcox 

Structural Health Monitoring

1–13

© The Author(s) 2022



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/14759217221113240

[journals.sagepub.com/home/shm](https://journals.sagepub.com/home/shm)

## Abstract

Previously, a nonlinear autoregressive network with exogenous input (NARX) demonstrated an excellent performance, far outperforming an established method in optimal baseline subtraction, for defect detection in guided wave signals. The principle is to train a NARX network on defect-free guided wave signals to obtain a filter that predicts the next point from the previous points in the signal. The trained network is then applied to new measurement and the output subtracted from the measurement to reveal the presence of defect responses. However, as shown in this paper, the performance of the previous NARX implementation lacks robustness; it is highly dependent on the initialisation of the network and detection performance sometimes improves and then worsens over the course of training. It is shown that this is due to the previous NARX implementation only making predictions one point ahead. Subsequently, it is shown that multi-step prediction using a newly proposed NARX structure creates a more robust training procedure, by enhancing the correlation between the training loss metric and the defect detection performance. The physical significance of the network structure is explored, allowing a simple hyperparameter tuning strategy to be used for determining the optimal structure. The overall detection performance of NARX is also improved by multi-step prediction, and this is demonstrated on defect responses at different times as well as on data from different sensor pairs, revealing the generalisability of this method.

## Keywords

Baseline subtraction, defect detection, time-series data prediction, guided wave, machine learning

## Introduction

Guided wave testing is a commonly used non-destructive evaluation (non-destructive evaluation NDE) technique often applied to structural health monitoring (SHM), as in comparison to bulk wave inspections, it requires fewer sensors and less operator time in collecting test signals. This is because waves from permanently attached sensors can propagate over long distances in waveguides (e.g. plate-like structures and pipes).<sup>1</sup> The signals obtained from a network of sensors can be processed to form images to locate any defects in the inspected regions.<sup>2–4</sup> In many structures, measured guided wave signals are dominated by the responses from structural features making the detection of responses from defects challenging. The general solution to this is based on recording one or more baseline signals when the structure is in a defect-free state, to which subsequent measurements can be compared. This forms the basis of a defect detection strategy for long-term monitoring called baseline signal subtraction.<sup>5–7</sup> However, as the inspection may span

multiple years during the lifetime of a structure, the changing environmental and operational conditions (EOCs), especially temperature, will distort the baseline signals of a defect-free structure.<sup>8</sup> These signals and their changes are complex, making it hard to distinguish real defect responses from those arising from changes in EOCs.

Various compensation techniques have been developed to reduce the effect of temperature on the baseline signal. This includes baseline signal stretching,<sup>9–11</sup> optimal baseline selection (OBS),<sup>9,12</sup> a combination of the two,<sup>13</sup> as well as more recent developments compensating for a range of effects besides the change in velocity.<sup>14,15</sup> OBS, which selects the best matching baseline signal from a pool of signals, then amplitude stretched

Department of Mechanical Engineering, University of Bristol, Bristol, UK

### Corresponding author:

Xin L Tu, Department of Mechanical Engineering, University of Bristol, University Walk, Beacon House, Queens Road, Bristol BS8 1QU, UK.  
Email: [xt16846@bristol.ac.uk](mailto:xt16846@bristol.ac.uk)

and time stretched,<sup>13</sup> has been compared to another data-driven approach, nonlinear autoregressive network with exogenous input (NARX), which is a machine learning (ML) method. The latter has shown superior defect detection performance.<sup>16</sup> Previous work with NARX<sup>16</sup> processed data spanning 8 years from a sparse array of permanently attached sensors on a steel tank. The network was trained using defect-free signals recorded when the structure was assumed to be pristine, and predicted baseline signals with matching EOCs when given synthetic defect test signals. The degradation of the tank over several years was assessed, and the defect detection performance of the network was further validated by introducing a physical anomaly to the tank. Other ML methods have been investigated for anomaly classification<sup>17,18</sup>; however, they require carefully designed example defect data for training.

The NARX network is an autoregressive neural network containing a single hidden layer that predicts a future point in a time-domain guided wave signal based on previous points (history) in the signal. The history points are repeated in a specific way to feed a wider input layer for the NARX network. This paper aims to improve the training robustness and physical interpretability of the network structure, by understanding how changing the length of history input, the width of the

performance. Section 7 generalises the use of this network structure to different datasets.

## NARX structure

In the current paper, NARX is implemented as a fully connected neural network with one hidden layer, which takes  $n_h$  history values from guided wave signals to predict the  $n_f$ th step into the future, where  $n_h$  and  $n_f$  are defined as history length and future prediction length, respectively. The output of the network  $\hat{s}[n+n_f]$  is obtained by mapping a function for all the  $n_h+n_f-1$  values prior to the prediction point:

$$\hat{s}[n+n_f] = \hat{f}\{s[n-n_h+1], s[n-n_h+2], \dots, s[n-1], s[n], \hat{s}[n+1], \dots, \hat{s}[n+n_f-1]\}, \quad (1)$$

where  $n$  denotes the current step in time,  $s$  denotes known values from data and  $\hat{s}$  denotes predicted values.

A flow chart illustrating the NARX implementation on time-series data is shown in Figure 1(a).  $n_h$  values are taken from the data (Stage 1) and split to  $n_l$  signals of length  $n_u$ , each delayed by one step (Stage 2), where  $n_u+n_l-1=n_h$ . This is then concatenated to form the input vector  $\mathbf{s}_{in}$  consisting of  $n_l \times n_u$  values:

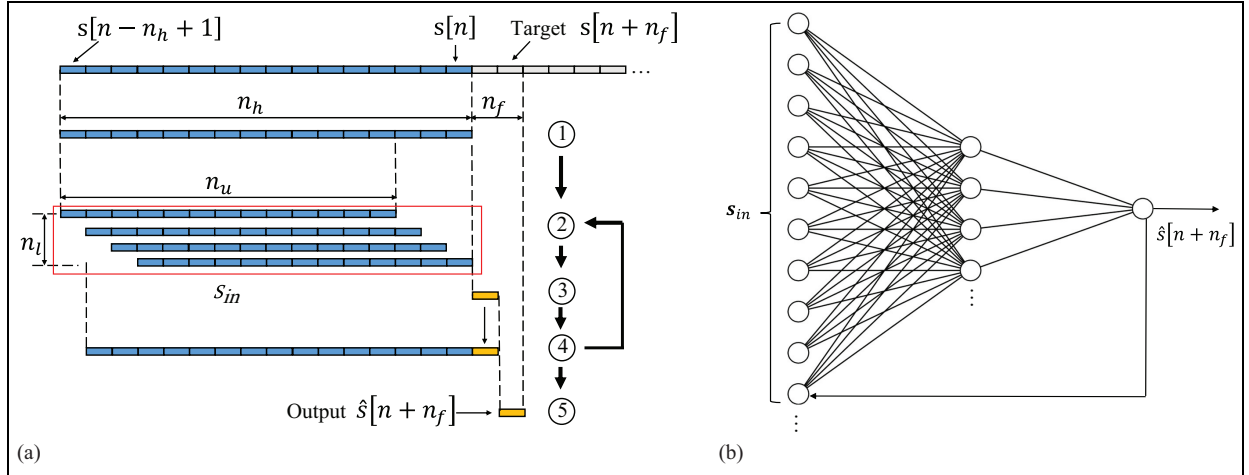
$$\mathbf{s}_{in} = \left\{ \begin{array}{cccc} s[n+1-n_l-n_u], & s[n+1-n_l-(n_u-1)], & \dots, & s[n+1-n_l], \\ s[n+1-(n_l-1)-n_u], & s[n+1-(n_l-1)-(n_u-1)], & \dots, & s[n+1-(n_l-1)] \\ \dots, & \dots, & \dots, & \dots \\ s[n+1-n_u], & s[n+1-(n_u-1)], & \dots, & s[n] \end{array} \right\} \quad (2)$$

input layer and the number of future steps to predict ahead, affects its detection performance, using the same experimental data as previous work.<sup>16</sup> Subsequently, it aims to show the training procedure and detection performance of the network is robust to different initialisations of weights, defect responses at different times and data from different sensor pairs, revealing the potential of generalising such a network for similar defect detection problems in SHM and NDE.

The paper is organised as follows. Section 2 introduces the NARX structure, and Section 3 the data used for training, validating and testing NARX. In Section 4, the primary drawback of the previously implemented NARX is presented. Section 5 proposes an optimal network structure using a new approach to hyperparameter tuning. Section 6 shows that the new structure overcomes the drawback of the previous implementation, providing more consistent, better defect detection

Therefore  $n_l \times n_u$  is referred to as the input layer width. In this way,  $n_u$  in the current work merges the  $n_u$  and  $n_y$  in previous work,<sup>16</sup> and  $\mathbf{s}_{in}$  merges the two inputs which were termed  $u$  and  $y$ . This improves clarity but does not affect operation of the network in any way.

$\mathbf{s}_{in}$  is passed through the network shown in Figure 1(b), and a predicted value for the next step is given (Stage 3). For single-step ahead prediction, the value at Stage 3 is the final output, whereas for multi-step prediction, this value joins the previous  $n_h-1$  values, which are regrouped and passed through the network by iterating over Stages 2–4. This process is equivalent to taking  $n_h$  values starting at different steps ( $s[n-n_h+1], s[n-n_h+2], \dots$ ) from the right-hand side of Equation (1) to form a new  $\mathbf{s}_{in}$  in each loop. In this case, single-step ahead prediction (SP mode of NARX<sup>19</sup>) is a special case of multi-step prediction (P mode<sup>19</sup>), where the feedback loop shown in Figure 1(b) is not used.



**Figure 1.** (a) Example flow chart of NARX implementation on time-series data. Blue indicates known values taken from data, and yellow indicates predicted values. (b) Schematic drawing of a NARX network. NARX: nonlinear autoregressive network with exogenous input.

The total number of weights and biases of the network is given by  $[n_l \cdot n_u + 2]n_m + 1$ , where  $n_m$  is the number of hidden units. In the current paper,  $n_m = \text{ceil}(\sqrt{(n_u + n_y)n_l + n_{ol}} + 10)$  as established in various studies,<sup>20,21</sup> where  $(n_u + n_y)n_l$  is the number of nodes in the input layer and  $n_{ol}$  is the number of nodes in the output layer.

Current work migrated the NARX network from Matlab to Tensorflow, as the latter provides greater capacity for processing large datasets and more flexibility in adapting network structure through lower level coding. This migration necessitates the use of a different optimiser, Adam, which is commonplace in Tensorflow, rather than Levenberg–Marquardt, the default for NARX in Matlab. The default learning rate of 0.001 was found to yield best performance. A batch size of 256 was found to be optimal among a range tested (from 32 to the training sample size). The network is trained to minimise the mean squared error (MSE) loss metric between  $s[n + n_f]$  and  $\hat{s}[n + n_f]$ .

## Data for training, validating and testing NARX

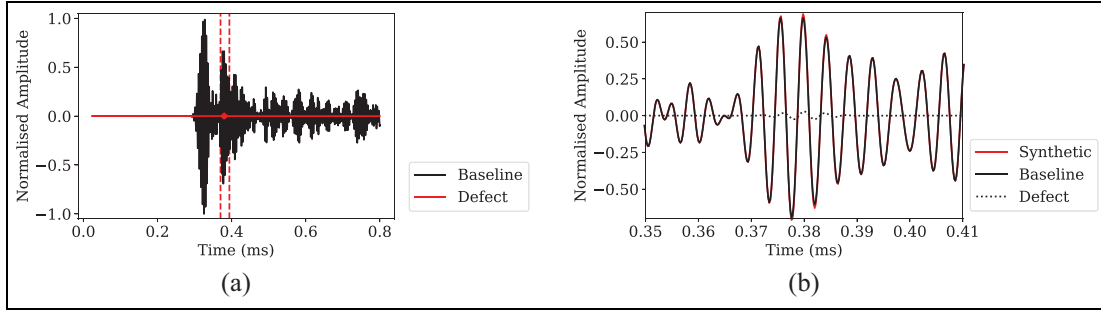
The experimental data consist of guided wave signals obtained from nine piezoelectric disc sensors permanently installed on a steel water tank, and monitored from 2012 to 2020, following the previous work.<sup>16</sup> The data from the years 2012 and 2013, which are the immediate measurements after sensor installation, are assumed to represent the tank's pristine condition for the purposes of subsequent monitoring. The current work first investigated the NARX network in detail with data from one sensor pair: transmit on Sensor 1 and receive on Sensor 8. The training data are

comprised of 20 signals chosen from the 2012 set over a relatively even distribution of EOCs. The validation data and defect-free test data are comprised of 5 and 200 signals from 2013. Later, more training data are added at random and data using different sensor pairs are also investigated.

The signals transmitted by the sensors are chirp excited,<sup>22</sup> and deconvolved to a five-cycle Hanning windowed toneburst with a centre frequency of 250 kHz,<sup>16</sup> and propagate primarily as  $S_0$  mode. The data were sampled at 5 MHz, which gives 20 points per cycle. Each signal has a duration of 0.8 ms, and hence consists of 4000 points in time. Removing 120 points of electrical crosstalk at time zero leaves 3880 valid points. The training, validation and test samples were regrouped from each set of original signals to form inputs of  $n_h$  history length. Therefore, each dataset has a sample size of  $[\#signals \times 3880 - (n_h + n_f - 1)]$ .

Artificial defect reflections, which are scaled and delayed tonebursts, were added to the defect-free test signals to form synthetic signals. Only one such defect response is added to each signal. This model is thought to be more conservative than the real measurements, because it ignores subsequent echoes and shadowing, which are likely to be bigger than the direct reflection from the defect. The defect responses were scaled to the time of the first arrival signal, referred to as the Syn1 method in previous work,<sup>16</sup> to simulate the beam spreading effect, and were multiplied by another scale factor, referred to as severity (given in dB),<sup>16</sup> to represent the size of the defect response. Therefore, the amplitude  $a$  of the defect response is given by<sup>16</sup>:

$$a = \frac{2}{ct} 10^{(\beta/20)}, \quad (3)$$



**Figure 2.** (a) Baseline signal and added defect signal of amplitude  $-30$  dB. Red dashed line indicates the location of the defect in time. (b) Zoom-in of synthetic data at defect location.

where  $c$  is the group velocity of the guided wave,  $t$  is the arrival time of the defect response and  $\beta$  is the severity of the defect. The current work initially looks at defects with severity of  $-30$  dB occurring at  $t = 0.38$  ms, to distinguish the detection performances of different networks using the largest signal to noise ratio (SNR) level not detected reliably. Later the network is also tested with defects occurring later in time, which have smaller amplitudes than at  $t = 0.38$  ms. An illustration of an example defect-free test signal with added defect is shown Figure 2(a) and (b).

A standard criterion is set to measure the performance of networks, a detailed description of which can be found in previous work.<sup>16</sup> The receiver operating characteristic (ROC) curve of each defect is computed using probability of false alarm and probability of detection (POD) at different detection thresholds. If the maximum amplitude of a defect-free residual signal exceeds a threshold, then it indicates a false alarm; if the maximum amplitude of a residual signal with defect exceeds a threshold, then it counts towards POD. The area under curve (AUC) is calculated for each ROC. Higher AUC indicates the detection performance for that defect is better. The lower limit of AUC is 0.5, equivalent to random guessing.

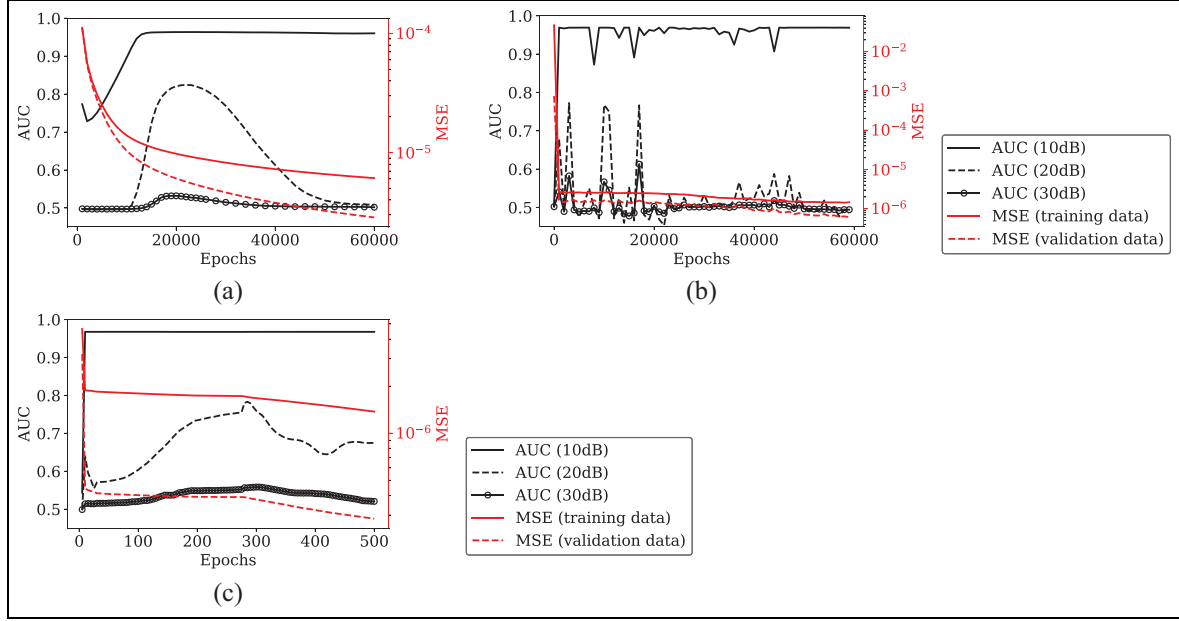
### Shortcomings of single-step prediction

To understand the limitations of the existing approach, various training algorithms, learning rates, batch sizes and two different ML environments (Matlab following previous work<sup>16</sup> and Tensorflow) were investigated for the previous NARX structure:  $n_u = 21$ ,  $n_l = 4$  and  $n_f = 1$ . Some representative results are presented in Figure 3, which show the MSE of training and validation data versus epochs as well as the AUC computed for different defect severities as training progresses. In all cases, it can be seen that the MSE for both training and validation data decreases monotonically during training. However, the MSE is not representative of the

detection performance of the network. Detection performance is described by the AUC and this can be seen to fluctuate up and down during the training process. This means that while there are windows of high AUC during training, these do not always appear at the same stage or have the same width. Therefore, it is difficult to define a generalisable training protocol to obtain networks with consistently good detection performance. It is hypothesised that when predicting one step into the future, the network simply extrapolates the next point from the previous few points rather than actually learning the characteristics of the pristine response. Therefore, even when tested with defect signals, the network still manages to predict the next step well enough to obscure the defect response in the residual signal. At that stage, the network is considered to be overfitted, for example, from Epoch 30,000 onwards in Figure 3(a). The overfitting cannot be detected with the usual criteria, shown by an increase in validation loss while the training loss continues to decrease, because the problem is set in a way that the AUC does not track monotonically with MSE – the network can produce a small residual signal regardless of whether it is defect free or with defects.

Increasing the correlation between MSE and AUC will improve the robustness of the training procedure, which is the main aim of this work. Validation loss is used as an indicator of how well the model will perform at its intended task: detecting defects in the test set. Although ultimately the detection performance is indicated by AUC, this is not suitable to be used as a loss metric – not only would this ramp up the computational burden, it would also require defect signals to be present in training data. This leads to difficult questions such as how big the defect should be, at what time should they occur, and are they representative of all possible defects of interest, etc, and bypasses the aim of defect-free model training.

Subsequently, multi-step prediction is investigated with the aim of reducing the disparity between the



**Figure 3.** AUC, and MSE for both training and validation data evaluated over training history of single-step prediction NARX for training algorithm: (a) GDX in Matlab, (b) Adam in Python, and (c) Levenberg–Marquardt (default for NARX) in Matlab. AUC: area under curve; MSE: mean squared error; NARX: nonlinear autoregressive network with exogenous input; GDX: gradient descent with adaptive learning rate and momentum training.

training loss metric, MSE, and the performance metric, AUC. It is believed that by predicting more steps into the future, the network will be unable to simply extrapolate forward as a way to reduce MSE between predicted and true signals. This means that when applied to a signal containing a defect response, the network should be unable to predict the defect response from data before it is visible in its input.

### Hyperparameter study for multi-step prediction

This section searches for a network structure that can achieve excellent detection performance, which will subsequently enable the investigation of the AUC and MSE correlation for training robustness in the following sections. Therefore, a hyperparameter tuning strategy assisted by physical reasoning is presented, showing an efficient way to find an optimal structure without needing to test all possible combinations. Essentially, the performance of the NARX network is influenced by the following factors: how much history information is input to the network ( $n_h$ ), how large the network is ( $n_u \times n_l$ ), how many steps into the future the network is predicting ( $n_f$ ) and how much training data is used. All these factors are interrelated, and it is hypothesised that increased input history length would require a larger network to process (more input signals and network complexity), which subsequently gives the

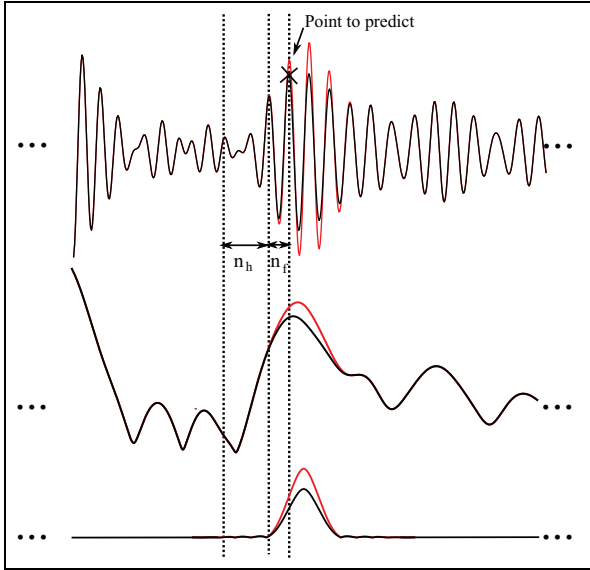
network the potential to predict further ahead in time. Therefore, best possible detection performances are compared for each of those factors, which are used as indicators for optimising the hyperparameters.

For the used digital time domain signals, it is more physically significant to describe  $n_h$ ,  $n_u$ ,  $n_l$  and  $n_f$  in terms of their relative scale to a toneburst in the actual signal rather than numbers of time steps. Therefore, the duration of a toneburst  $T_l$  is used, which is given by:

$$T_l = \frac{n}{f_c}, \quad (4)$$

where  $n$  is the number of cycles in a toneburst and  $f_c$  is the centre frequency of the signal. Figure 4 shows how the history length  $n_h$  and future prediction length  $n_f$  of a multi-step prediction NARX are related to a typical signal. The top graph illustrates when predicting  $\geq 1$  cycle ( $0.2T_l$ ) ahead, the input to the network would still be the same as the baseline signal, whereas the network is required to predict a value where the defect response shows a local maximum. In this case, the predicted value would follow the baseline signal (black line), and subtracting that from the test data (red line) would return a residual more representative of the defect response. The bottom graph shows a simpler representation of the principle using envelopes of windowed tonebursts at the defect location. When  $n_f$  is too small, the difference between the predicted baseline and the





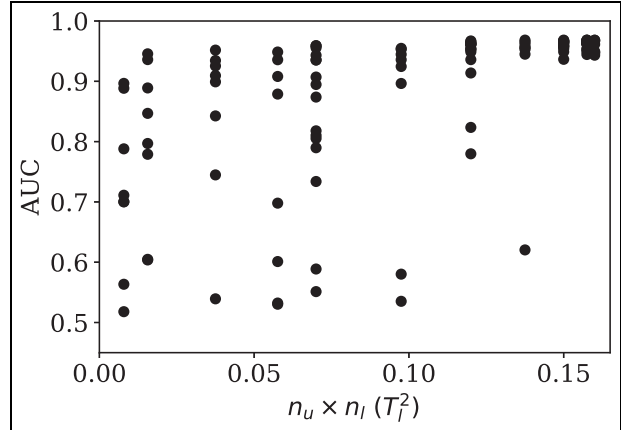
**Figure 4.** (Top) Extraction of data from an example baseline guided wave signal (black line) to be input to NARX network, and an example defect signal (red line) showing deviation from baseline signal where the defect occurs. (Middle) Signal envelopes of the example baseline and synthetic defect signal in the top graph. (Bottom) A simplified representation using only the envelopes of the windowed tonebursts at defect location. NARX: nonlinear autoregressive network with exogenous input.

test data would not be significant enough to allow robust defect detection.

### Changing input layer width

This section aims to find the optimal combination of  $n_u$  and  $n_l$  for a certain  $n_h$ . This means for a given history length  $n_h$ , how complex the network should be to achieve the best performance, as the input layer width  $n_u \times n_l$  determines the number of hidden units and subsequently the total number of neurons in the network. Ideally, the network is set to predict 0.5 of a toneburst ahead ( $n_f = 0.5T_l$ ) so that the output will still be predicting the defect-free response right up to the position of the defect peak. However, given  $n_f = 0.01T_l$  in previous work, a more modest value  $n_f = 0.2T_l$  (1 ultrasonic cycle ahead) is tested first and subsequent sections will explore extending this.

To predict one cycle ahead,  $n_h$  must be large enough to provide sufficient history information.  $n_h \approx 4n_f$  is chosen to start with, which means that four cycles of history length is input to the network, and the effect of increasing  $n_h$  will be investigated later in this paper. The study was performed using different values of  $n_u$  ranging from  $0.01T_l$  to  $0.7T_l$  and  $n_l = 0.8T_l - n_u$ , and the networks trained with eight different initialisations for each combination of  $n_u$  and  $n_l$  to understand the variance resulting from random initialisations of



**Figure 5.** Best performance (AUC) the network can achieve throughout the training at different complexities with given history length  $n_h = 0.8T_l$ , when  $n_f = 0.2T_l$ . AUC: area under curve.

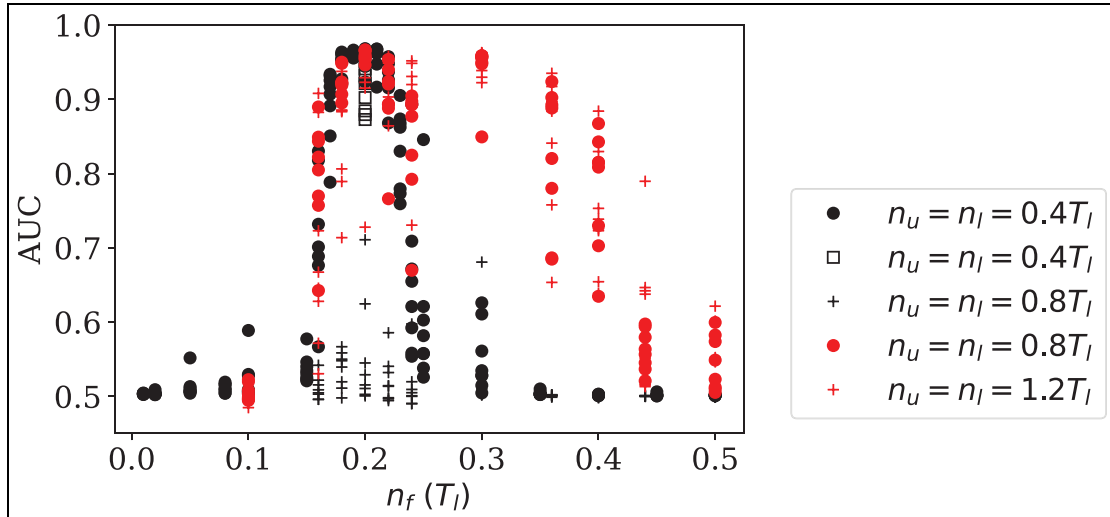
weights. All networks were trained over 8000 epochs to allow sufficient convergence on MSE.

Figure 5 plots the best AUC the network can achieve for detecting  $-30$  dB defects in training history versus  $n_u \times n_l$ . It is found that for  $n_u \times n_l < 0.1375T_l^2$ , high detection performance is not achieved reliably, with the scatter in AUC having a standard deviation of 0.130 for  $n_u \times n_l = 0.008T_l^2$ . The performance of the network becomes progressively more consistent as it increases in size. The standard deviation for AUC is only 0.013 for  $n_u \times n_l = 0.16T_l^2$ . As a result,  $n_u = n_l$  is used in subsequent studies to ensure the optimal performance of the network, as this is the hard limit of how complex the single-layer NARX-based network can be for a given history length,  $n_h$ .

### Changing future prediction length

The history length  $n_h = 0.79T_l$  with maximum input layer width  $n_u = n_l = 0.4T_l$  is then tested with different future prediction lengths,  $n_f$ , up to predicting half a toneburst ahead ( $0.5T_l$ ). Figure 6 shows that detection for the  $-30$  dB defect is only achieved with  $0.15T_l < n_f < 0.25T_l$  (black dots), although it is hypothesised that the network should also perform well past this point up to  $n_f = 0.5T_l$ . When the defect signal is at its maximum at  $n_f = 0.5T_l$ , a greater residual could be obtained after subtraction, making the defect more detectable.

Before more tests were carried out to see whether the network only performs best with the structure  $\{n_u, n_l, n_f\} = \{0.4T_l, 0.4T_l, 0.2T_l\}$  in this scenario, the data were down-sampled to reduce the computation load. This is best illustrated by considering predicting half a toneburst ahead, the network should at least



**Figure 6.** Best performance (AUC) the network can achieve throughout the training when predicting different points ahead. Results in black are from training data with 20 signals, results in red are from training data with 60 signals. Apart from results on black dots, all other results are obtained from down-sampled data. AUC: area under curve.

have  $n_u = n_l = 1T_l$ , resulting in 111 hidden units and 1,121,323 total parameters, which took more than 72 h to train on a graphics processing unit (GPU) facilitated high-performance computer. Halving the sampling rate to 2.5 MHz results in 10 points per cycle yielded similar performance as shown in Figure 6 by black squares. This means the information content in the signals remains the same after down-sampling. But it reduced the computation time significantly because the number of parameters in the network was reduced by 84% and the number of feedback loops in the network, the training, validation and test data points the network has to go through were halved when the network was scaled down.

### Changing history length

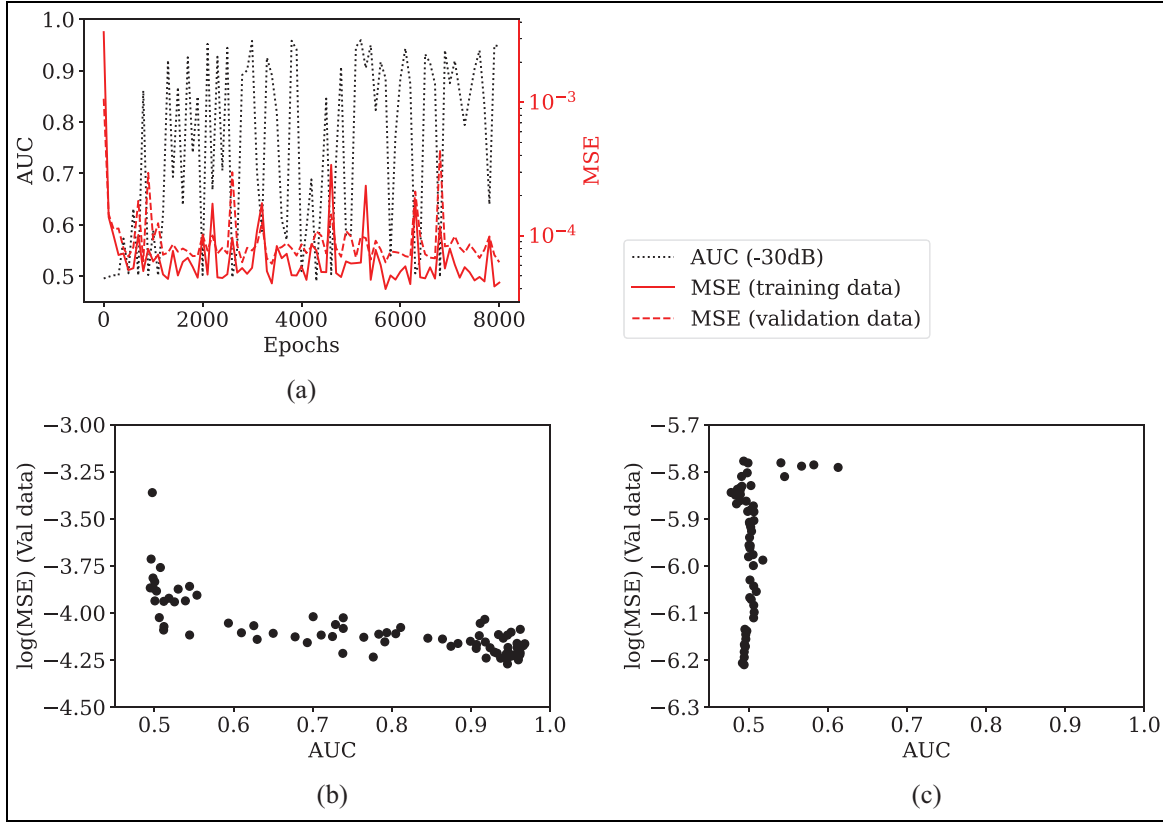
Networks with increased history lengths are also assessed to determine the performance over different future prediction lengths to see whether the range for peak performances is extended this way. Both  $n_u$  and  $n_l$  are increased by the same amount to achieve maximum network complexity. Black crosses in Figure 6 show the detection performance when doubling the history length ( $n_h = 1.6T_l$ ), which are worse than those of  $n_h = 0.8T_l$  in previous section. It is hypothesised that the deteriorated performances are due to insufficient training data for the comparatively larger networks resulting from the increased history length. Consequently, to ensure training sample size is not a limiting factor, an extra 40 signals selected at random are added to the training data. This effectively triples the training sample

size while the history length doubles. The performances from those are shown by the red dots in Figure 6. It can be seen that the lower limit for good detection is still set at  $n_f = 0.15T_l$ ; however, with more training data, future prediction length can be increased to  $n_f = 0.4T_l$  before performance degrades. Increasing the history length further to  $n_h = 2.39T_l$  with the current 60-signal training data does not show notable improvements on the detection performance (red crosses in Figure 6).

On the other hand, increasing  $n_u$  alone would result in unbalanced  $n_u$  and  $n_l$ , which leads to sub-optimal network structures as the detection performances would show less consistency compared to those with  $n_u = n_l$ . Therefore, results on those tests are not shown.

To summarise, the peak performance around  $n_f = 0.2T_l$  is shown to be indefinite, meaning it is a function of training data size and history length. It is also shown that achieving good performance does not require the network to be very large, as the number of parameters in a network is many fewer in down-sampled tests. It is reasonable to assume that given enough training data, the network would be able to predict 0.5 toneburst ahead, although it is quite computationally expensive in this scenario. In this way, when supplied with adequate training data and setting  $n_u = n_l$  for maximum network complexity, NARX, as a multi-layer perceptron, works as a universal approximator.<sup>23,24</sup> Further work is required to determine the optimal training set size for a given  $n_h$  or the other way around. The training set size beyond which the performance stops improving is yet to be explored.





**Figure 7.** (a) AUC, and MSE for training, validation and test data evaluated at every 100 epochs over the training history of a multi-step prediction NARX with optimal structure  $\{n_u, n_l, n_f\} = \{0.4T_l, 0.4T_l, 0.2T_l\}$ . (b)  $\log(\text{MSE})$  of validation data scattered against AUC for multi-step prediction NARX, values taken from (a). (c)  $\log(\text{MSE})$  of validation data scattered against AUC for single-step prediction NARX, values taken from Figure 3(b).

AUC: area under curve; MSE: mean squared error; NARX: nonlinear autoregressive network with exogenous input.

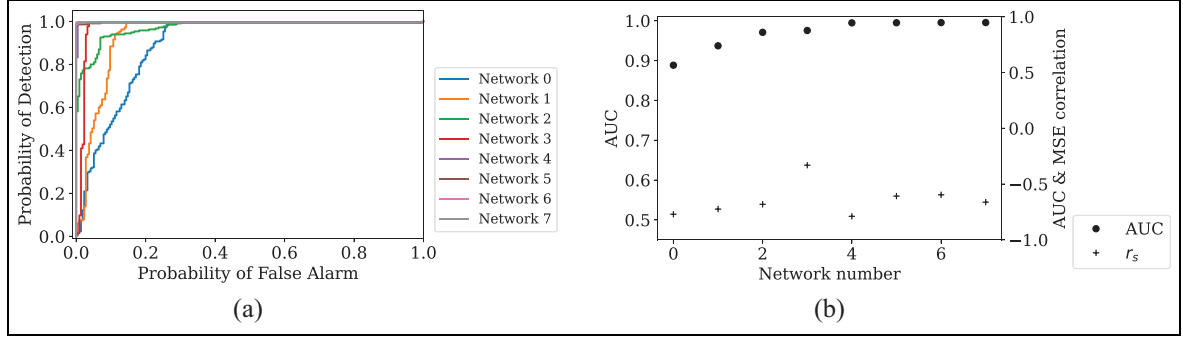
Due to the presence of some abnormal measurements in the 2013 test data (4 out of 200 signals tested for sensor pair 1–8), the maximum AUC only reaches 0.97 in Figure 5 and 6. Those abnormal signals can be clearly identified from the test data, as their amplitudes are one order of magnitude lower than others. It is believed that these are a result of occasional poor electrical contacts in the relay-based multiplexors used to acquire experimental data from the array of sensors. The network can still distinguish the defects from noise in those residuals; however, a different threshold is required as the residuals are significantly lower than others, limiting the maximum AUC to below unity. The abnormal signals are omitted from the test data in subsequent sections to present the true detection performance of the network.

### Repeatability of proposed NARX structure

The training metric, MSE of validation data, can be used as a good proxy for the desired performance

metric, AUC, provided there is a strong correlation between the two. Therefore, this correlation is investigated using a training history of the proposed optimal structure of multi-step prediction NARX. AUC, and MSE for training, validation and test data are evaluated every 100 epochs, giving 80 checkpoints throughout training, which are plotted as training curves in Figure 7(a). The MSE of validation data is then plotted against AUC on a log scale in Figure 7(b), showing a clear monotonic trend – as MSE decreases and AUC improves. In contrast,  $\log(\text{MSE})$  against AUC for single-step prediction in Figure 7(c) (with values from training graph in Figure 3(b)) shows a lack of correlation, which explains why training lacks robustness.

The Spearman correlation coefficient,  $r_s$ , which shows the strength and direction of the monotonic relationship between two variables,<sup>25</sup> is computed to quantify the correlation between AUC and MSE. Therefore, it eliminates the effect of outliers, for example, at early stages of training, when test MSE falls rapidly but AUC has not improved yet, or at late stages of training, when test MSE continue to decrease but AUC reached its ceiling. It is given by:



**Figure 8.** (a) ROC for detecting  $-30$  dB defects by best performing networks trained with eight different initialisations. Networks saved at the point of lowest validation loss (MSE of validation data). Networks are ordered by detection performance from low to high. (b) Left axis: calculated AUC from the ROC curves. Right axis: Spearman correlation coefficient,  $r_s$ , for AUC and MSE of validation data.

AUC: area under curve; MSE: mean squared error; ROC: receiver operating characteristic.

$$r_s = 1 - \frac{6 \sum D_i^2}{N(N^2 - 1)}, \quad (5)$$

where  $D_i$  is the difference between the  $X$  and  $Y$  ranks for the  $i$ th case,  $N = 80$  is the sample size.<sup>25</sup> The  $r_s$  value for  $\log(\text{MSE})$  and AUC is  $-0.72$  for Figure 7(b) and  $-0.030$  for Figure 7(c), thus confirming that the former exhibits a strong negative correlation, while the previous training approach is effectively random.

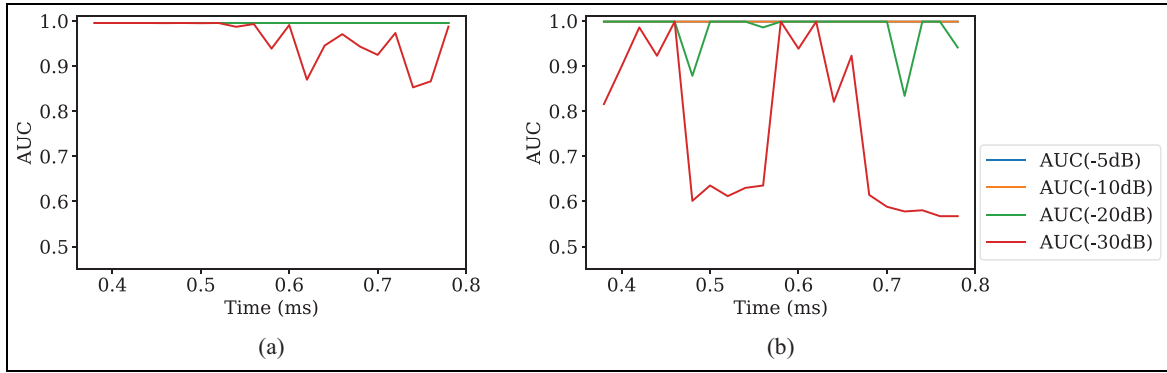
On this basis, the point of lowest MSE of validation data in training history in general leads to a network with good detection performance. Therefore, the multi-step prediction network with the optimal structure found in the previous section,  $\{n_u, n_l, n_f\} = \{0.4T_l, 0.4T_l, 0.2T_l\}$  is trained eight times with different initialisations, and saved at the point with lowest validation loss within 8000 epochs of training, using a custom callback function. The detection performance is assessed when detecting defects with a severity of  $-30$  dB occurring at  $t = 0.38$  ms, because this defect amplitude is the largest one that cannot be detected reliably by all networks. The ROC curves are shown in Figure 8(a), and are subsequently used to compute the AUCs in Figure 8(b), which have a mean of  $\mu = 0.89$ , and a standard deviation of  $\sigma = 0.061$ . The  $r_s$  values for AUC and MSE in 8b for each of the networks have a mean of  $\mu = -0.65$  and a standard deviation of  $\sigma = 0.13$ , showing a good correlation between the training metric and the performance metric for each. It can be seen that training the current NARX network in this way (saving the network weights at lowest MSE) consistently returns good detection performance, and there is a high probability of obtaining a perfect detector (when  $\text{AUC} = 1$ ) for  $-30$  dB defects.

The multi-step prediction NARX network also demonstrates improved detection performance when tested with defects occurring at different times compared

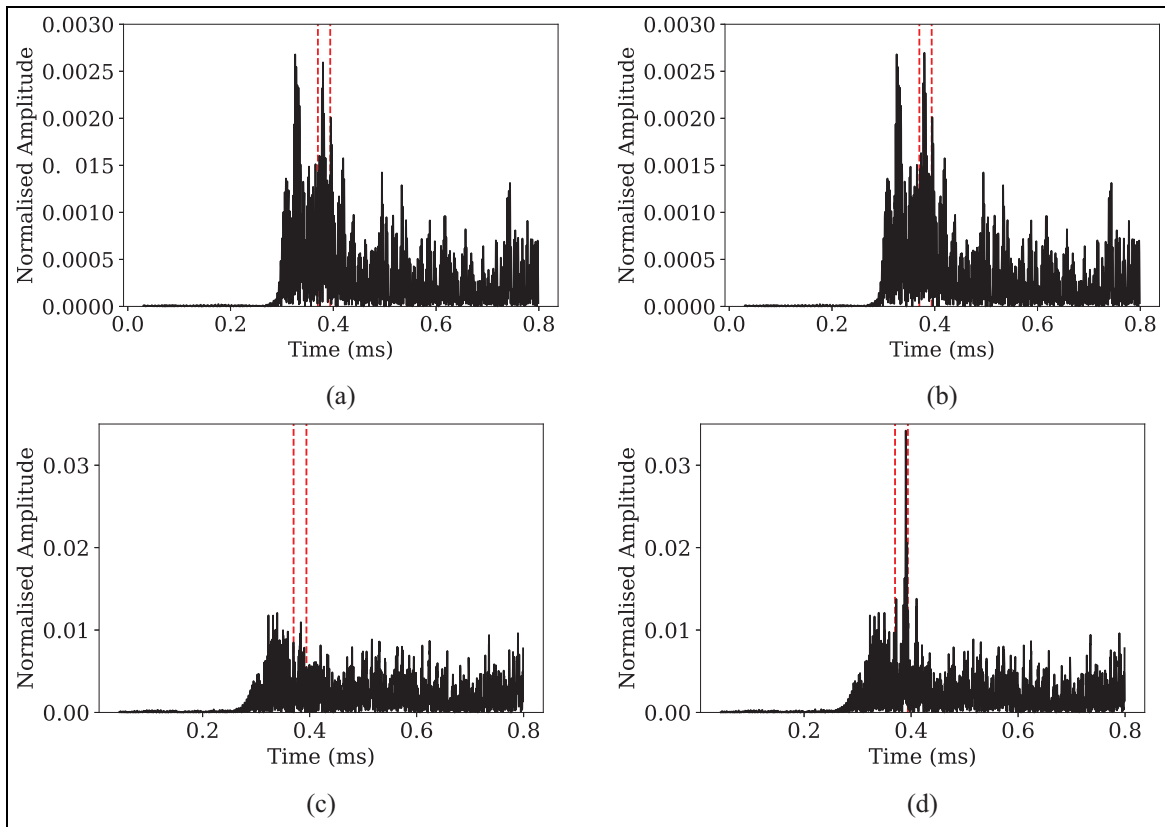
to the previously implemented single-step prediction network (Figure 9(a) and (b)). From the perspective of signal processing, single-step prediction gives unevenly distributed noise in the residual signal when subtracted from defect-free test data, as shown in Figure 10(a) and (b). Multi-step prediction with the optimal hyperparameters found in Section 5, on the other hand, gives more evenly distributed residual signal (Figure 10(c)). They can also maintain a good SNR when subtracted from test data with defects (Figure 10(d)), even though the average residual is much higher than that of single-step prediction networks (Figure 10(a) and (b)). The residual amplitude at the defect location is also more physically representative, as the defect in this case has an amplitude of  $0.025$  above the average residual. Comparing to its true amplitude of  $0.027$ , this means defect responses are passed through the network largely unaffected. In other words, multi-step prediction with the optimal structure enhances the correlation between AUC and MSE and achieves good detection by reducing the occurrence of large values in the residual.

## Generalisability of proposed NARX structure

The proposed optimal network structure,  $\{n_u, n_l, n_f\} = \{0.4T_l, 0.4T_l, 0.2T_l\}$ , is applied to signals from other sensor pairs to evaluate its performance across different datasets. For each sensor pair, (transmitter–receiver) 1–2, 1–3, 1–4, 1–5, 2–4 and 2–5, the network is trained using five different initialisations, with a training sample comprised of 60 signals from the 2012 set chosen at random. Artificial defects are added to the defect-free test data from the 2013 set using the same method, Syn1, as in Section 2. The defects have a severity of  $-30$  dB. Their normalised amplitudes are kept the



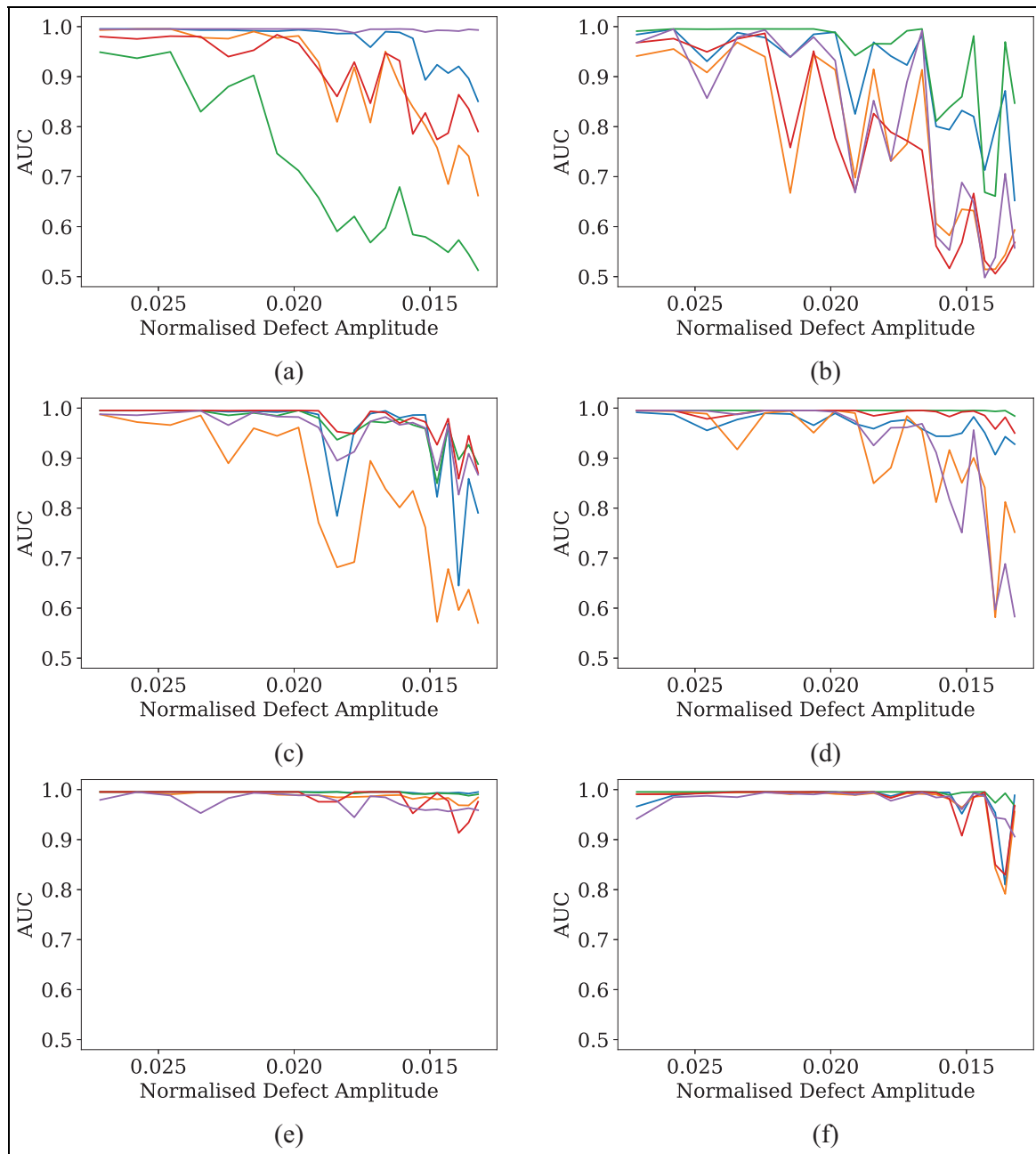
**Figure 9.** (a) Detection performance of multi-step prediction NARX with optimal structure  $\{n_u, n_l, n_f\} = \{0.4T_l, 0.4T_l, 0.2T_l\}$  for defects occurring at different times. (b) Detection performance of previously implemented single-step prediction NARX for defects occurring at different times, reproduced from previous work,<sup>16</sup> but with anomalous readings excluded to allow a fair comparison with the current results.  
 AUC: area under curve; NARX: nonlinear autoregressive network with exogenous input.



**Figure 10.** Residual signal from single-step prediction NARX (a) without defect (b) and with defect. Residual signal from multi-step prediction NARX with optimal structure (c) without defect (d) and with defect.  
 NARX: nonlinear autoregressive network with exogenous input.

same as the defects in the time location test in Figure 9 (the peak of the test-free data is normalised to 1). In this case, the defects are added to different time locations for different sensor pairs due to different first

arrival time of the signals. The results are presented in Figure 11. At the first instance ( $-30$  dB with normalised amplitude = 0.027), all networks can detect the defects from all sensor pairs with AUCs well above 0.9



**Figure 11.** Defect detection performance of proposed optimal NARX structure trained and tested with data from (a) sensor pair 1–2, (b) sensor pair 1–3, (c) sensor pair 1–4, (d) sensor pair 1–5, (e) sensor pair 2–4 and (f) sensor pair 2–5. Defects appearing later in time are detected with smaller amplitudes due to beam spreading, so the x-axis ‘Normalised Defect Amplitude’ decreases from left to right. It follows the same direction of time as in Figure 9. NARX: nonlinear autoregressive network with exogenous input.

and little variation. The detection performance of the networks is better and more consistent when the defects have greater amplitude ( $>0.025$ ); it degrades as the defect amplitude decreases. In general, the results show

that the network is readily applicable to more sensor pairs in addition to the one studied in detail in the current work. Future work will involve generalising the network to other NDE applications through transfer

learning or by adapting the network structure using the hyperparameter tuning method presented earlier.

## Conclusion

In this paper, the single-step prediction NARX network is found to have inconsistent defect detection performance, because the loss metric for training, MSE, is not a reliable indicator of the performance metric, AUC. By predicting multiple steps ahead, the NARX network is able to achieve a strong inverse correlation between MSE and AUC. This enables a robust training procedure to be defined, and the best performing network can be reliably selected from training history at the point of lowest MSE. The overall detection performance of multi-step prediction is also seen to be better than single-step prediction. This method is generalisable for data from different sensor pairs.

It is shown that by looking at the physical significance of the network with regard to the guided wave signals, a simple hyperparameter tuning for the optimal structure can be performed. This involves starting with a reasonable value of input history length based on a physically representative future prediction length, and then testing different future prediction lengths using the maximum network complexity for the given input history length. The input history length and the training set size should be increased when predicting further ahead.

Current work also reveals some questions that are worth future investigation, for example, how to quantitatively assess the effect of adding more input history length and training data, and where the upper limits of performance lie.

## Acknowledgements

The authors would like to thank Kangwei Wang for providing Matlab code and data for training and testing NARX networks in previous work.

Current work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - <http://www.bristol.ac.uk/acrc/>.

## Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.


## Funding


The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The work reported is a continuation of a pilot project (grant no. 100374) originally supported by Lloyd's Register


Foundation and the Alan Turing Institute Data-Centric Engineering Programme.

## ORCID iDs

Xin L Tu  <https://orcid.org/0000-0001-9115-3683>

Richard J Pyle  <https://orcid.org/0000-0002-5236-7467>

Anthony J Croxford  <https://orcid.org/0000-0003-1377-2694>

Paul D Wilcox  <https://orcid.org/0000-0002-8569-8975>

## References

1. Mitra M and Gopalakrishnan S. Guided wave based structural health monitoring: a review. *Smart Mater Struct* 2016; 25: 053001.
2. Hay T, Royer R, Gao H, et al. A comparison of embedded sensor lamb wave ultrasonic tomography approaches for material loss detection. *Smart Mater Struct* 2006; 15(4): 946.
3. Yan F, Royer RL Jr and Rose JL. Ultrasonic guided wave imaging techniques in structural health monitoring. *J Intell Mater Syst Struct* 2010; 21(3): 377–384.
4. Putkis O and Croxford AJ. Continuous baseline growth and monitoring for guided wave SHM. *Smart Mater Struct* 2013; 22(5): 055029.
5. Michaels JE and Michaels TE. Detection of structural damage from the local temporal coherence of diffuse ultrasonic signals. *IEEE Trans Ultrason Ferroelectr Freq Control* 2005; 52(10): 1769–1782.
6. Worden K, Farrar CR, Manson G, et al. The fundamental axioms of structural health monitoring. *Proc R Soc A: Math Phys Eng Sci* 2007; 463(2082): 1639–1664.
7. Clarke T, Simonetti F, Rokhlin S, et al. Evaluation of the temperature stability of a low-frequency a0 mode transducer developed for shm applications. In: *34th Annual Review of Progress in Quantitative Nondestructive Evaluation 2008*, Golden (Colorado), USA, vol. 975, pp. 910–917. American Institute of Physics.
8. Konstantinidis G, Wilcox PD and Drinkwater BW. An investigation into the temperature stability of a guided wave structural health monitoring system using permanently attached sensors. *IEEE Sens J* 2007; 7(5): 905–912.
9. Lu Y and Michaels JE. A methodology for structural health monitoring with diffuse ultrasonic waves in the presence of temperature variations. *Ultrasonics* 2005; 43(9): 717–731.
10. Croxford AJ, Wilcox PD, Konstantinidis G, et al. Strategies for overcoming the effect of temperature on guided wave structural health monitoring. In *Health monitoring of structural and biological systems* 2007, 2007, vol. 6532, p. 65321T. International Society for Optics and Photonics.
11. Harley JB and Moura JM. Scale transform signal processing for optimal ultrasonic temperature compensation. *IEEE Trans Ultrason Ferroelectr Freq Control* 2012; 59(10): 2226–2236.
12. Konstantinidis G, Drinkwater BW and Wilcox PD. The temperature stability of guided wave structural health monitoring systems. *Smart Mater Struct* 2006; 15(4): 967.

13. Croxford A, Moll J, Wilcox P, et al. Efficient temperature compensation strategies for guided wave structural health monitoring. *Ultrasonics* 2010; 50: 517–528.
14. Mariani S, Heinlein S and Cawley P. Location specific temperature compensation of guided wave signals in structural health monitoring. *IEEE Trans Ultrason Ferroelectr Freq Control* 2019; 67(1): 146–157.
15. Mariani S, Heinlein S and Cawley P. Compensation for temperature-dependent phase and velocity of guided wave signals in baseline subtraction for structural health monitoring. *Struct Health Monit* 2020; 19(1): 26–47.
16. Wang K, Zhang J, Shen Y, et al. Defect detection in guided wave signals using nonlinear autoregressive exogenous method. *Struct Health Monit* 2022; 21(3): 1012–1030.
17. Ying Y, Garrett JH Jr, Oppenheim IJ, et al. Toward data-driven structural health monitoring: application of machine learning and signal processing to damage detection. *J Comput Civil Eng* 2013; 27(6): 667–680.
18. Hesser DF, Kocur GK and Markert B. Active source localization in wave guides based on machine learning. *Ultrasonics* 2020; 106: 106144.
19. Menezes JJ and Barreto G. Long-term time series prediction with the narx network: An empirical evaluation. *Neurocomputing* 2008; 71: 3335–3343.
20. Keles D, Scelle J, Paraschiv F, et al. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. *Appl Energy* 2016; 162: 218–230.
21. Marcjasz G, Uniejewski B and Weron R. On the importance of the long-term seasonal component in day-ahead electricity price forecasting with narx neural networks. *Int J Forecasting* 2019; 35: 1520–1532.
22. Michaels JE, Lee SJ, Croxford AJ, et al. Chirp excitation of ultrasonic guided waves. *Ultrasonics* 2013; 53(1): 265–270.
23. Wang H and Song G. Innovative narx recurrent neural network model for ultra-thin shape memory alloy wire. *Neurocomputing* 2014; 134: 289–295.
24. Hornik K, Stinchcombe M and White H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989; 2: 359–366.
25. Myers J, Well A and Lorch JR. *Research design and statistical analysis (3rd ed.)*. New York: Routledge, 2013.