Evaluation of Temperature As an Exogenous Input to Modelling the Structural Health of a Material

Peter Fox
School of Engineering Mathematics & Technology
University of Bristol
Bristol, England
ox18249@bristol.ac.uk

Tung Lam Ha
School of Engineering Mathematics & Technology
University of Bristol
Bristol, England
qq24422@bristol.ac.uk

Aidan McKillop
School of Engineering Mathematics & Technology
University of Bristol
Bristol, England
vu21904@bristol.ac.uk

Wenxiang Zhang
School of Engineering Mathematics & Technology
University of Bristol
Bristol, England
tn24943@bristol.ac.uk

***Abstract*—One of the primary challenges in guided wave-based structural health monitoring (SHM) is that environmental variability, particularly temperature changes, can mask or mimic defect signals. This study investigates the integration of temperature as an exogenous input within machine learning models to improve defect detection accuracy. Using eight years of experimental data from a steel storage tank, NARX-based neural networks, LSTM regressors, and a temperature-conditioned autoencoder were evaluated. Results show that while conventional models gain minimal benefit from temperature inclusion, the autoencoder significantly enhances predictive accuracy, achieving an R^2 of 0.87 and substantially reducing reconstruction error. By leveraging temperature-aware architectures, this work advances the development of intelligent monitoring systems capable of reliable defect detection in complex, real-world environments.**

I. Introduction

Structural health monitoring (SHM) involves the implementation of damage detection strategies for components across a range of mechanical infrastructure systems [1]. It measures the structural response of key parameters under environmental or operational conditions and commonly employs sensing techniques such as acoustic emission and ultrasonics to evaluate the health status of structures and detect potential damage at an early stage [2], [3]. Among available techniques, guided wave structural health monitoring (GWSHM) has emerged as a crucial non-destructive evaluation method, particularly for large-scale infrastructure such as steel storage tanks [4]. By analyzing high-frequency ultrasonic signals propagating through sensor networks, GWSHM enables large-area defect detection with a minimal number of sensor deployments [5].

A significant challenge in GWSHM lies in distinguishing true defect signals from environmental noise caused by operating conditions, particularly temperature variations. As Croxford et al. [6] note, temperature fluctuations can alter wave velocity and attenuation, leading to a significant increase in baseline drift in traditional subtraction methods. The Nonlinear Autoregressive with eXogenous inputs (NARX) model has shown potential in learning dynamic patterns from healthy-state signals to predict baseline responses. Nevertheless, Wang et al. [5] found that when exogenous variables such as temperature are neglected, the performance of NARX models deteriorates in time-varying environments. This study aims to explore the impact of incorporating temperature as an exogenous input variable into differing NARX models for guided-wave defect detection, with the goal of enhancing the quality of residual signals and the reliability of detection.

Project repository available at: https://github.com/EMATM0050-2024/dsmp-2024-group30

II. Literature Review

A. NARX Without Exogenous Input

One widely adopted approach for identifying defects in structural health monitoring is NARX; a recurrent neural network which can model dynamic systems by predicting future values based on past observations with exogenous inputs [7]. In SHM, NARX models learn the typical behavior of guided wave signals in a healthy structure. Wang et al. [5] applied this approach by training a NARX model on a data set of guided wave signals collected when a structure was in pristine condition. The network learned how these signals typically behaved under different environmental conditions, without requiring the temperature to be explicitly provided as an input. By subtracting the predicted baseline from the actual measured signal, they were able to isolate unexpected changes that could indicate the presence of defects. This method proved particularly effective in overcoming challenges posed by varying environmental conditions, where traditional techniques such as optimal baseline selection (OBS) [6] often struggle.

However, the initial single-step prediction approach showed inconsistencies, as the network tended to focus on short-term patterns. To address this, Tu et al. [7] introduced multistep

prediction to force the network to capture broader signal behavior and improve both stability and detection performance. Both studies demonstrated that well-tuned NARX models can effectively enhance defect detection even without temperature input.

B. Incorporation of Exogenous Data in Structural Health Monitoring

In the field of structural health monitoring, the accurate detection of structural damage is significantly challenged by environmental factors, particularly variations in temperature. To address this, studies have focused on integrating exogenous variables, primarily temperature, to mitigate environmental impacts. A notable example is Croxford et al. [8], who proposed the Optimal Baseline Selection, whereby a database of baseline signals is collected and used alongside interpolation to match experienced conditions with an appropriate value to subtract. This approach, however, requires extensive collection of baseline data and may result in imperfect subtraction if the temperature interpolation is inaccurate.

An alternative compensation method proposed by Mariani et al. [9] involves adjusting the phase velocity of the toneburst excitation signal to account for thermal effects, otherwise known as signal stretching.

C. Other Considered Model Types

Several machine learning models have been applied to the analysis of SHM, providing alternatives to the NARX-based approach undertaken by Wang et al. [5]. In a comprehensive review of 40 published models, Sattarifar et al. [10] found that 79% employed supervised techniques, with principal component analysis (PCA) accounting for 41% of those investigated, followed by continuous wavelet transforms (CWT) and then NARX-based models. Among unsupervised methods, k-means clustering was the most frequently implemented.

Other recent studies have explored using generative approaches to model training. Khurjekar et al. [11] trained an ensemble of variational autoencoders (VAE) on simulated structural data and then tested the trained model on temperature-varying experimental data. The model demonstrated robust defect detection across thermal variations and outperformed comparative deep learning architectures.

III. Methodology

A. Comparison of temperature vs non-temperature input

As identified during the literature review, multiple approaches to modelling have been proposed to account for the effects of temperature in SHM, particularly to minimise amplitude distortions following signal subtraction.

Although compensation-based strategies are currently the predominant means of addressing temperature variation in SHM, this study proposes an alternative: incorporating temperature as an exogenous input to predictive models. This approach enables models to internally learn the impact of temperature on resulting traces, and potentially capture and adjust for seasonal patterns relevant for long-term deployment scenarios in practice.

To evaluate this hypothesis, models were trained and compared under three conditions: one without the inclusion of temperature, one with raw temperature values, and one with temperature along with trend values to account for long-term fluctuations.

These trend values included temperature delta with respect to the previous trace, 24 hour rolling average temperature and deviation from monthly average temperature. By including this information, it was predicted that wider temperature change behaviour and seasonality across samples could be factored when training the model, leading to a more accurate representation of temperature, and thus a lower residual.

Model performance would then be assessed to determine (i) whether the inclusion of temperature improves signal accuracy and (ii) which model configuration demonstrated superior overall performance.

B. Modelling a NARX-based Neural Network

Two lightweight feedforward neural network models were developed based on the NARX framework [5] to evaluate the impact of temperature features on defect detection in guided wave signals. Both models retain the core logic of the original NARX approach, following a pipeline of feature normalisation, model training, and residual analysis, and employ a synthetic defect generation method similar to that described by Croxford et al. [5]. Statistical features, specifically energy and kurtosis, were extracted from the guided wave signals to characterise waveform behavior. Synthetic defect signals were generated based on physical parameters such as wave velocity, sensor spacing, and defect severity, and superimposed onto pristine signals to simulate damage conditions. The resulting datasets were normalised and labeled to distinguish between pristine and defect-injected signals.

The original NARX architecture, based on recursive output feedback and tapped delay lines, was simplified by adopting fully connected feedforward neural networks. This modification reduced computational complexity while maintaining the ability to learn dynamic features. Hyperparameters, such as the embedding dimension ($n_u = 20$), were initialised using optimal values determined via mutual information and Cao's method [5]. The number of neurons in the hidden layer (n_m) was computed based on the size of the input and optimal lag using:

$$n_m = \left\lceil \sqrt{(n_u + 1) \cdot n_l + 1} + 10 \right\rceil \tag{1}$$

where n_l represents the optimal lag obtained through mutual information analysis, and n_u is the predefined input delay. The models were trained using the mean squared error (MSE) loss function:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

Two model variants were investigated. The first, denoted as NARX-Lite, omits exogenous variables such as temperature but retains the extraction and normalisation of signal-derived statistical features. By processing time-series data alone, this model emphasises learning temporal characteristics inherent to the guided waveforms. The second variant, NARX-Temp, extends the NARX-Lite structure by incorporating temperature data as a static feature, combining environmental variables with waveform statistics.

While temperature is not dynamically modeled, its inclusion enables the assessment of environmental influences on defect detection performance and provides a foundation for future extensions integrating temporal and environmental factors.

Post-training, residuals between predicted and actual signals were analyzed, and a Receiver Operating Characteristic (ROC) curve was generated to evaluate the detection capability of each model.

C. Modelling a NARX-based LSTM regression

A sliding window autoregressive stacked Long Short-Term Memory (LSTM) network was then implemented to predict guided wave signals for defect detection.

The input to the LSTM consisted of sequences formed using a sliding window approach. For each prediction step, the model was provided with $n = 10$ previous timesteps of the guided wave signal along with corresponding exogenous temperature values. This structure allowed the LSTM to capture short-term dynamics while integrating environmental effects. The input matrix $\mathbf{X}_t$ at timestep t is defined as:

$$\mathbf{X}_t = \begin{bmatrix} x_{t-n} & e_{t-n} \\ x_{t-n+1} & e_{t-n+1} \\ \vdots & \vdots \\ x_{t-1} & e_{t-1} \end{bmatrix} \quad \rightarrow \quad \hat{x}_t \tag{3}$$

where x represents the signal amplitude and e represents the temperature at each timestep. The LSTM was trained to predict $\hat{x}_t$, the signal at timestep t, using a one-step-ahead forecasting strategy across the entire dataset.

Regarding model structure, an LSTM layer with 64 units and tanh activation was initially added. A relatively high unit count was selected to capture complex, long-range temporal dependencies across the input space. The tanh activation function was chosen for its ability to model both positive and negative signal deviations effectively. A Dropout layer of rate 0.3 was applied after the first LSTM layer to mitigate overfitting, providing a balance between retaining sufficient model capacity and preventing the memorization of training sequences.

A second LSTM layer comprised of 32 units with tanh activation was then added. Reducing the number of units encouraged the model to distil learned temporal features ahead of prediction, reducing the risk of overfitting. A second Dropout layer of rate 0.3 was inserted between the LSTM layers and the final output, which was then followed by a fully connected dense output layer producing 4999 outputs, corresponding to the complete predicted waveform.

The model was compiled using the Adam optimiser with a learning rate of 0.001. Mean squared error (MSE) was used as the loss function to directly penalise large prediction deviations. This configuration was selected to balance model complexity against the risk of overfitting, taking into account the high dimensionality of the waveform data and the modest size of the training set.

After training, the LSTM-generated predictions $\hat{x}_t$ were compared to the actual measured signals x_t to compute residuals:

$$r_t = x_t - \hat{x}_t \tag{4}$$

These residuals were then analysed to detect deviations indicative of structural defects. Further spectral analysis using Fast Fourier Transform (FFT) was performed on the residuals to identify frequency-domain anomalies associated with damage.

Additionaly, an encoder-decoder autoencoder with temperature conditioning was employed as an unsupervised method to reconstruct our guided wave data. This method stems from multiple iterations of testing with similarly structured LSTM and CNN-based models. Accurately predicting future waveforms requires capturing both historical wave dynamics and concurrent temperature effects.

This autoencoder architecture is designed to incorporate these aspects: the encoder compresses historical waveform data through a series of causal convolutional layers combined with maximum pooling operations. The encoder's convolutional filters are structured to capture temporal dependencies and the underlying wave patterns, summarising these into a condensed latent vector. Current temperature data is also embedded into a low-dimensional latent representation via dense layers. These two latent spaces, waveforms and temperature, are concatenated to generate a combined vector, which is then expanded and reshaped to serve as the initial input for the decoder. The decoder employs upsampling techniques and convolutional refinements to reconstruct full future waves.

D. Evaluation Approach

To evaluate the performance of the various models, a set of well-established, statistically significant metrics was employed. For the classification model, five indicators were adopted, namely ROC-AUC, accuracy rate, precision rate, recall rate, and F1 score, to evaluate performance. In this study, due to the category imbalance between the test set and the training set, the recall rate, F1, and ROC-AUC were primarily emphasized to avoid the influence of the accuracy rate, which focuses only on the detection ability of samples from a single category.

For the regression model, the coefficient of determination (R^2) was used to quantify the proportion of variance in the true waveforms captured by the model, providing an intuitive measure of overall fit. Mean Squared Error was included

to heavily penalize significant errors, which is crucial when outliers carry important meaning. Mean Absolute Error (MAE) was used to provide a clear sense of the average error magnitude within a model. The Median Absolute Error (MedAE) was reported to capture a robust, outlier-resistant summary of typical deviation. For the unsupervised methods, where a predicted wave substitutes missing data, a representative test trace was overlaid against its ground truth to visually assess alignment. Finally, all residuals across test waves were aggregated into a histogram to inspect their distribution.

IV. Data Description / Preparation

A. Description Of Data And Pair 3 Selection

The research data were derived from sparse-array piezoelectric sensors installed on a steel water tank. An automatic data acquisition system was employed to collect data, capable of obtaining the time-domain response signals of all sensor pairs in each measurement cycle. Following de-chirping processing, these signals were recorded as datasets. Beginning in 2012, the system performed 76,680 measurements, with the experiment suspended in July 2016 and restarted in July 2020. A total of 76,680 signal datasets from 12 sensor pairs were ultimately obtained. Each data file (.mat) contained three components: trace data with 6,000 signal points, time data recording the collection timestamp, and "Data1" metadata mapping sensor and defect positions.

As illustrated in Figure 1, analysis of defect positions via the "Data1" metadata revealed a high coincidence with the path between Sensor 1 and Sensor 4 (i.e., pair 3). In terms of spatial correlation, this sensor pair more directly and accurately captured guided-wave signal features related to defects. When selecting a single dataset for preliminary analysis, choosing pair 3 based on this strong correlation enhanced the pertinence and accuracy of defect detection. From a signal propagation perspective, the positions of Sensor 1 and Sensor 4 rendered guided-wave signals more significantly influenced by defects during propagation, making signal change features easier to capture.

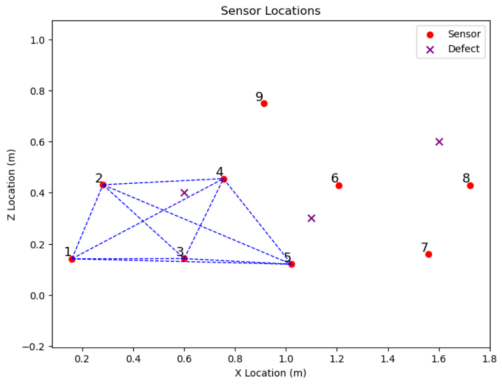

Fig. 1. Sensor and defect position mapping

B. Trace Subtraction Analysis

In order to conduct subtraction analysis, the real and imaginary parts of the signal were separated to simplify the process and focus on amplitude variations that were more indicative of structural changes. Since guided wave signals are complex-valued, isolating the real part provides a clearer representation of physical wave behavior relevant for defect detection.

The subtraction was performed by calculating the difference between consecutive signal acquisitions using:

$$\Delta x_i = x_i - x_{i-1} \tag{5}$$

where x_i is the signal at the current timestamp and x_{i-1} is the previous signal. This approach highlights anomalies by emphasising sudden changes over time.

When a defect is introduced, the subtraction result exhibits significant fluctuations around the main wave packet, as shown in Figure 2. In contrast, when no defect is present, the difference remains close to zero, forming a nearly flat line.

It was also observed that higher sampling intervals lead to increased fluctuations due to temperature variations affecting wave propagation. This demonstrates the influence of environmental factors, reinforcing the need to account for temperature changes in structural health monitoring.

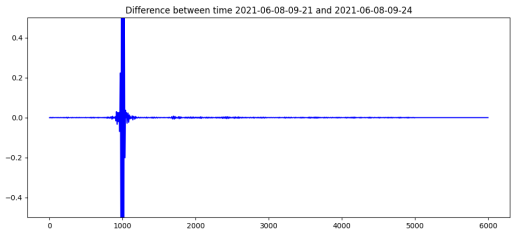

Fig. 2. Subtraction results: Significant fluctuation indicating defect introduction

C. Trace Sample Variation

Before any signal processing and sampling, the periods during which data from the chosen sensor pairing were available were first quantified. Data was available over nine and a half years, from January 2012 to June 2021. A dense record existed from early 2012 through late 2015, typically obtaining between 1200 and 2500 samples per month, with six months of complete inactivity from September 2013 to February 2014. Coverage from November 2015 to April 2016 was sparse, with fewer than 500 total samples. These numbers increased again from May to July 2016 (approximately 3500 total samples). A lengthy period of inactivity followed from July 2016 until mid-2020, during which there were between 200 and 700 traces per month in the final six months of 2020. Finally, on June 8, 2021, between 9 and 11 a.m., a total of 34 traces were recorded; this was later determined to correspond to the simulation of defects occurring at that time.

To discover wave trends over time, the root-mean-square (RMS) amplitudes of the signals were plotted to visualise any changes. From early 2012 through early 2015, the RMS

remained steady at approximately (0.065 ± 0.005), with minor fluctuations throughout. However, around mid-2015, the RMS dropped dramatically by approximately 50% and remained at this lower level until the recordings ceased (mid-2016 to mid-2020). After this prolonged period of inactivity, the RMS amplitude of the sensor readings no longer behaved uniformly, exhibiting extreme variation in values over short periods. As such, it was determined that training models using data collected before this large RMS drop-off would be most prudent.

D. Correlation of Trace and Weather Data

To justify the inclusion of exogenous inputs in modelling, weather data associated with the trace acquisitions was obtained and prepared. Hourly temperature and humidity data for Bristol- where the experimental data was collected – were retrieved from Open Meteo's free weather API for the full duration of 2012, during which time the traces could be truly considered pristine and defect-free.

Each trace in the 12-month period was paired with the nearest corresponding weather record within a 15-minute window of the acquisition timestamp. To ensure comparability between features, all waveform and weather-related columns were normalised using min–max scaling to a [0, 1] range.

Analysis was then completed to determine if correlation and subsequent causality could be determined between trace responses and the weather data collected.

E. Data Sampling

To ensure a manageable yet representative subset of trace data was used during model training and testing, an evaluation into the implementation of two differing approaches to trace data sampling was undertaken, with results compared against evaluation metrics.

1) 2012/13 Data: To create a consistent and representative training and testing corpus, traces collected throughout 2012 were used for training, while traces from 2013 were reserved for testing. Both time periods fall within the first two years of experimental operation and are considered to contain pristine, defect-free samples. The rationale behind this temporal split was to ensure that both training and testing datasets captured a full year of seasonal variation and temperature fluctuations, enabling the models to learn and predict temperature-dependent behaviours more effectively. To ensure comparability, both datasets underwent identical processes of data preparation and normalisation.

For each sample set, the full year's trace data were loaded and aligned to corresponding hourly weather data retrieved from Open Meteo, matched within a 15-minute window of each trace acquisition. Additional columns indicating the month, hour, day, and a discretised temperature bin were appended to the resulting data frames. These features were subsequently used to create stratified samples of approximately 2000 traces for training and 500 traces for testing.

Following stratified sampling, monthly temperature bin distributions were assessed. In instances where specific months exhibited low or missing trace counts, synthetic samples were generated using scikit-learn's resampling function to balance the distributions. The overall temperature distributions were then validated to confirm coverage across the seasonal spectrum, and the datasets were finalised for use in model training. After sampling and refinement, the final datasets consisted of 1770 training samples and 465 testing samples.

2) First 5.5 years: A stratified, seasonally balanced sampling procedure was applied across the first five and a half years of data availability (excluding only 2020 and 2021). Choosing this timeline as the "undamaged set" for model training was intended to enable the models to learn any new defects; however, it may not allow the models to revisit historical data (from before mid-2016) and distinguish older defects. A sample size of approximately 10% of all points within the selected timeframe was selected, with targets distributed evenly across seasons to maintain meteorological seasonality throughout the sample set. This approach ensured the elimination of bias towards specific times of the year, allowing for an accurate comparison of differential temperature conditions.

V. Results and Discussion

A. Correlation Analysis

The relationship between guided wave trace values and environmental conditions was investigated to determine the suitability of temperature as an exogenous input in modelling.

A correlation matrix of raw trace data to both temperature and humidity showed a moderate positive correlation with temperature (r = 0.66), and a weak negative correlation with humidity (r = -0.33). To further investigate lagged relationships, cross-correlation functions between trace data and both temperature and humidity were plotted.

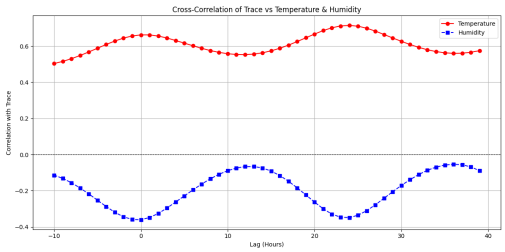

Fig. 3. Cross-correlation of trace values with temperature and humidity.

As illustrated in Figure 3, the correlation between temperature and trace signals peaked at a lag of approximately 24 hours ($r \approx 0.7$), indicating a potentially meaningful temporal relationship. In contrast, the strongest correlation for humidity ($r \approx -0.36$) remained at zero lag, suggesting a limited temporal effect. Based on these results, temperature was selected as the single exogenous input to be used in modelling.

After confirmation of moderate correlation between trace and temperature, Granger Causality Tests were performed over a 30-hour lagged window on a random sample of 500 traces to ensure computational efficiency and reduce the risk of

overfitting. Given the moderate sample size, the SSR-based F-test was employed as the principal statistical test due to its finite-sample properties. At a lag of 26 hours, the F-test indicated statistically significant Granger causality (F = 1.87, p = 0.0065), a finding that was further supported by the SSR-based chi-square ($\chi^2 = 54.76$, p = 0.0008) and likelihood ratio tests ($\chi^2 = 51.82$, p = 0.0019).

Based on these findings, the inclusion of temperature as an exogenous input in the modelling framework was deemed to be statistically justified.

B. Evaluation of Temperature Feature Representations

As described in section A of the methodology, three LSTM models were trained using the 2012 trace data: one with no exogenous inputs, one including only raw temperature, and one incorporating raw temperature along with three additional temperature trend features. All models were trained using a sliding window approach over ten full trace waveforms, with the model configuration described in Section F, "Unsupervised Model Evaluation". Performance was evaluated on 2013 test data, with results summarised in Table III.

TABLE I
COMPARISON OF MODEL PERFORMANCE WITH RESPECT TO TEMPERATURE FEATURES

Temperature Feature	R^2	MSE	MAE	MedAE
No Temperature	0.20287	0.00032	0.00992	0.00787
Raw Temperature	0.19747	0.00031	0.00988	0.00784
Raw Temperature & Trends	0.17863	0.00033	0.01029	0.00826

Based on the results in Table III, the model incorporating raw temperature and trend features performed worst across all evaluation metrics. Although its MSE value remained comparable to that of other models, its R^2 value was 12% lower than the no-temperature model, and the remaining metrics (MAE, MedAE) were 4-5% worse. These findings suggest that the addition of extra trends, while originally predicted to provide richer context information, may have introduced redundancy or noise. The high collinearity between the raw temperature and derived features, which reduces the unique information available to the model, could also have led to model overfitting, reducing overall predictive performance.

Comparing the models with and without raw temperature input, performance was similar across all metrics, with the temperature-included model outperforming in three out of four categories. Values of MAE and MedAE differed fractionally at 0.3% and 0.4% respectively, while the R^2 value for the no-temperature model was 3% higher. Overall R^2 values were low, likely due to the inherent nonlinear variability present in the trace waveforms. Finally, the model with temperature input achieved a 3% lower MSE compared to the no-temperature model, indicating better robustness to larger prediction errors.

These results suggest that including raw temperature as an exogenous input can provide modest improvement in LSTM model performance, primarily through reducing sensitivity to large prediction errors. However, the inclusion of additional temperature trend features did not yield any further benefit and may have introduced redundancy that degraded model accuracy.

C. Evaluation of Sampling Methods Used

As outlined in Section E, "Data Sampling," two strategies were proposed for model training and evaluation. The first trained on 2012 data and tested on 2013 data; the second used a 10% test-train split across the first 5.5 years of data collection, during which traces were assumed pristine. The training/testing splits were 1770/465 for the 2012–2013 set and 5228/2241 for the 5.5-year set, respectively.

To evaluate the sampling methods, two LSTM models were trained on the respective sets using a sliding window of ten full trace waveforms. Model configuration details are provided in Section F, "Unsupervised Model Evaluation," and performance results are summarised in Table II.

TABLE II
COMPARISON OF MODEL PERFORMANCE ACROSS SAMPLE SETS

Sample Span	R^2	MSE	MAE	MedAE
2012/2013 data	0.19747	0.00031	0.00988	0.00784
First 5.5 years	-0.00883	0.00092	0.01657	0.01394

The 5.5-year dataset showed significantly poorer performance, with a negative R^2 value indicating results worse than mean prediction. MSE was nearly three times higher than for the 2012–2013 set (0.00092 vs. 0.00031), and both MAE and MedAE approximately doubled, reflecting larger and more frequent outliers.

This degradation in model performance can be attributed to several factors. The extended multi-year dataset introduced considerable variability in environmental conditions and operational states, likely resulting in input distribution drift over time. Seasonal variations, equipment ageing, and changes in background noise levels may have compounded the problem. As both LSTM models were trained without explicit temporal correction for these shifts, the 5.5-year dataset trained model may have struggled to generalise across such evolving conditions, leading to higher error rates and unstable predictions. Sampling strategies that preserved the natural chronological order, such as the 2012–2013 split, more effectively maintained dataset consistency and temporal stability, supporting improved model generalisation and defect detection reliability.

Based on these findings, it was determined that all subsequent modelling would utilise the temporally separated 2012–2013 dataset to ensure greater consistency, minimise the impact of environmental drift, and enhance the reliability of model training and evaluation.

D. Supervised Model Evaluation

The supervised classification model combining a NARX structure and an LSTM layer demonstrated limited classification performance, as shown in Table III.

One contributing factor to the poor results was the significant sample imbalance within the test set. As the number

TABLE III
COMPARISON OF LIGHTWEIGHT FEEDFORWARD NEURAL NETWORK MODELS WITH AND WITHOUT TEMPERATURE

Temperature	Accuracy	Precision	Recall	F1	ROC-AUC
No Temperature	0.15047	0	0	0	0.80807
Temperature	0.35047	0.19747	0	0	0.74719

of defective samples greatly exceeded the number of non-defective samples, the model tended to predict the majority class, making it difficult to accurately identify non-defective instances.

Another challenge arose from the limited feature set used to represent signal characteristics. The extraction of only simple statistical features, such as energy and kurtosis, was insufficient to capture the complexity of real defect signals with varying degrees of damage. Consequently, the model learned based on incomplete or inaccurate feature representations, limiting its ability to distinguish between defective and normal signals. The low accuracy values (0.15047 without temperature and 0.35047 with temperature) shown in Table III reflect this limitation.

Furthermore, the limited similarity between simulated defect signals used during training and real defect signals encountered during testing reduced the model's generalisation capability. The ROC-AUC values (0.80807 without temperature and 0.74719 with temperature) indicate restricted discriminatory power, highlighting difficulties in adapting to real-world conditions.

E. Unsupervised Model Evaluation

1) LSTM: The LSTM models, both with and without raw temperature as an exogenous input, achieved modest performance across the evaluation metrics. The MAE and MedAE values were just below 0.01, indicating reasonably minimal deviation from the true waveforms. Additionally, MSE values of 0.00031 and 0.00032 for the models with and without temperature input, respectively, imply a relatively low absolute magnitude of prediction errors given the scale of the waveform features.

The R^2 values for both models were low at 0.197 and 0.203 for the raw temperature and no temperature models, respectively, indicating that approximately 20% of the variance in the predicted waveforms was explained relative to the actual signals. Although the positive values suggest that the models were able to learn some aspects of the underlying wave structure, the low overall R^2 indicates that a substantial proportion of the waveform variance remained unexplained.

Several factors may have contributed to the low R^2 scores. The input space was highly dimensional, with approximately 5000 features per time step, and the combination of a relatively modest training set size and a sliding window of 10 previous traces may have limited the model's ability to capture complex nonlinear dependencies. Moreover, the significant noise, variation, and long-range dependencies present in the input trace data may not have been optimally suited to a standard LSTM architecture without additional feature engineering or regularisation techniques.

Although the overall model performance was limited, the inclusion of raw temperature as an exogenous input provided a modest improvement across several evaluation metrics compared to models without temperature input.

2) Autoencoder: A comparison of LSTM, CNN, and autoencoder models for guided wave reconstruction is shown in Table IV. The LSTM model achieved an R^2 of 0.19747. To enhance performance, a CNN was tested as an intermediate step, resulting in a higher R^2 of 0.62203.

The final autoencoder model achieved the best results, with an R^2 of 0.87390 and the lowest errors across MSE, MAE, and Medae. The autoencoder was able to capture the waveform structure more effectively than previous models, benefiting from its ability to learn a compressed representation of the signal while incorporating temperature effects. These results demonstrate that the autoencoder approach provides a significant improvement in guided wave reconstruction accuracy over simpler LSTM or CNN-based methods.

TABLE IV
COMPARISON OF UNSUPERVISED MODEL ITERATIONS

Model	R^2	MSE	MAE	MedAE
LSTM	0.19747	0.00031	0.00988	0.00784
CNN	0.62203	0.00022	0.00801	0.00489
Autoencoder	0.87390	0.00013	0.00663	0.00316

Results also indicated that integrating temperature as an explicit conditioning variable significantly improves prediction accuracy and, thus, anomaly detection reliability. By explicitly embedding temperature information, the model can effectively compensate for environmental influences, thus enhancing overall defect detection. Comparative testing clearly demonstrated the necessity of including temperature, as the autoencoder without temperature input showed approximately a 13% reduction in R^2 score compared to a well-tuned model. This degradation in predictive accuracy was consistently observed across all iterations of unsupervised models tested and developed during this study.

Significantly larger reconstruction errors were observed in waveforms corresponding to structurally defective waveforms, illustrating the model's capability for anomaly detection. Figure 4 presents representative waveforms, showcasing apparent deviations between predicted and actual waveforms under defective scenarios.

Notwithstanding, several potential limitations with respect to the autoencoder model warranted discussion. Firstly, its performance depended heavily on the quality of the training dataset and how well it represented the structural health at the time of training. For the present dataset, insufficient coverage meant that, while the model performed exceptionally well in predicting waves from before mid-2016 (the large data gap), it was slightly less effective in predicting the most recent waves. Nevertheless, the model effectively identified significant (simulated) defects in the tank in 2021; however,

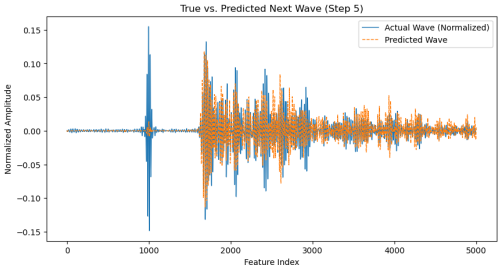

Fig. 4. Predicted Wave vs Actual Defective Wave

these defects altered the waveforms so dramatically that they could not be considered conclusive proof of concept.

Lastly, computational demands and the model's complexity may pose challenges, particularly for deployment in real-time monitoring scenarios. While the autoencoder-based model outperformed previous LSTM and CNN-based model iterations greatly, as shown in Table IV, both models were slightly less computationally intensive.

VI. Further Work and Improvement

Although the models outlined in this study show potential, particularly regarding temperature, several areas for further work have been identified. Future efforts should explore more comprehensive ways to utilise temperature information, such as modelling how the material's temperature changes over time and affects signal patterns. This could include capturing local variations in material temperature across the structure or examining how these shifts influence guided wave behaviour.

Model architectures could also be enhanced by placing greater emphasis on Transformer-based networks, which have demonstrated strong performance in capturing long-range temporal dependencies in sequential data [12]. Moreover, strategies such as domain adaptation or periodic retraining should be explored to better address changes in data distribution over time and ensure consistent model performance during extended operational use.

Ultimately, optimising the computational complexity of models, especially the autoencoder architecture, should be pursued to facilitate real-time deployment in embedded structural health monitoring systems. Integrating uncertainty quantification into predictions would further enhance the reliability and interpretability of future defect detection frameworks.

VII. Conclusion

This study highlights the importance of temperature as an exogenous variable in guided wave defect detection, based on a comparative analysis of NARX models with and without temperature inputs. The following key conclusions were drawn.

Incorporating raw temperature as an external input had a positive impact on model performance. Introducing temperature improved the LSTM model's predictive capabilities to a certain extent. However, the addition of temperature trend features did not enhance model performance and, in fact, led to a reduction in accuracy, likely due to the introduction of redundant information or noise. These results suggest that when utilising temperature information, the form of feature representation must be carefully selected to avoid negatively affecting model performance.

Furthermore, models trained on continuous-time data from 2012 to 2013 significantly outperformed those trained on the five-year mixed dataset. This finding emphasises that the stability of the time series is crucial for model generalisation. Distribution drift and seasonal differences introduced by random mixed sampling degraded model performance, whereas sampling strategies that preserve the temporal structure enabled a more reliable evaluation of temperature effects.

In terms of model comparison, the autoencoder demonstrated substantially better prediction performance than the LSTM regression model. By explicitly embedding temperature information, the autoencoder effectively compensated for environmental influences, thereby enhancing the reliability of anomaly detection. Nonetheless, certain limitations were observed, including a high dependency on the quality of the training dataset, increased computational requirements, and model complexity, which may present challenges for deployment in real-time monitoring scenarios.

References

[1] P. Kot, M. Muradov, M. Gkantou, G. S. Kamaris, K. Hashim, and D. Yeboah, "Recent Advancements in Non - Destructive Testing Techniques for Structural Health Monitoring," Applied Sciences, vol. 11, no. 6, p. 2750, 2021.

[2] F. J. Pallarés, M. Betti, G. Bartoli, et al., "Structural health monitoring (SHM) and Nondestructive testing (NDT) of slender masonry structures: A practical review," Construction and Building Materials, vol. 297, p. 123768, 2021.

[3] S. Hassani and U. Dackermann, "A systematic review of advanced sensor technologies for non - destructive testing and structural health monitoring," Sensors, vol. 23, no. 4, p. 2204, 2023.

[4] M. Castaings and B. Hosten, "Ultrasonic guided waves for health monitoring of high - pressure composite tanks," Ndt E International, vol. 41, no. 8, pp. 648 - 655, 2008.

[5] K. Wang, J. Zhang, Y. Shen, B. Karkera, A. J. Croxford, and P. D. Wilcox, "Defect detection in guided wave signals using nonlinear autoregressive exogenous method," in Structural Health Monitoring, vol. 21, no. 3, pp. 1012–1030, 2022.

[6] A. J. Croxford, P. D. Wilcox, B. W. Drinkwater, et al., "Strategies for guided - wave structural health monitoring," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 463, no. 2087, pp. 2961 - 2981, 2007.

[7] L. Tu, R. Pyle, A. J. Croxford, and P. D. Wilcox, "Potential and limitations of NARX for defect detection in guided wave signals," Structural Health Monitoring, vol. 21, no. 6, pp. 2040-2057, 2022.

[8] A. J. Croxford, J. Moll, P. D. Wilcox, et al., "Efficient temperature compensation strategies for guided wave structural health monitoring," Ultrasonics, vol. 50, no. 4 - 5, pp. 517 - 528, 2010.

[9] S. Mariani, S. Heinlein, and P. Cawley, "Compensation for temperature-dependent phase and velocity of guided wave signals in baseline subtraction for structural health monitoring," Ultrasonics, vol. 49, no. 10, pp. 765–773, Oct. 2009.

[10] A. Sattarifar and T. Nestorović, "Emergence of machine learning techniques in ultrasonic guided wave-based structural health monitoring: A narrative review," Int. J. Prognostics Health Manage., vol. 13, no. 1, pp. 1–21, May 2022.

[11] I. D. Khurjekar and J. B. Harley, "Closing the sim-to-real gap in guided wave damage detection with adversarial training of variational auto-encoders," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), May 2022, pp. 3588–3592.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.