DOI 10.5162/Ultrasonic2025/E4-a3

Investigation of a Deep Learning Methodology for Automatic Detection and Characterization of Crack-Type Defects in Ultrasonic Non-Destructive Testing

Patrick Scott Sheehan[1], Roberto Miorelli[1], Sébastien Robert[1], Sylvain Chatillon[2], and Bastien Chapuis[1]
[1]Université Paris-Saclay, CEA, List, 91191 Gif-sur-Yvette, France
[2]EDF, Direction Qualité Industrielle (DQI), 93206 Saint-Denis, France
roberto.miorelli@cea.fr

Abstract: The paper introduces a supervised simulation-based deep-learning pipeline for characterising welding defects with ultrasonic arrays in Non-Destructive Testing and Evaluation. Synthetic, multimodal TFM images generated with CIVA form the training and validation sets. The pipeline trains several state-of-the-art models using automated hyperparameter optimization. The trained models are then applied to experimental data collected under conditions mirroring the simulations. We compare regression performance and outline future directions.

Keywords: Ultrasonic testing, Deep-learning, Total Focusing Method, Simulation, Defect characterisation

Introduction

Within the ultrasonic Non-Destructive Testing and Evaluation (NDT&E) community, experimental datasets containing defects are limited since real defects in critical infrastructures are rare and most inspection results remain confidential. This scarcity hinders the effectiveness of Machine-Learning (ML) models, which rely on large and representative datasets to achieve reliable predictions.

This work employs images reconstructed by the Total Focusing Method (TFM) [1], one of today's most widely used ultrasonic imaging techniques in NDT&E. The TFM images are generated from Full Matrix Capture (FMC) data where each element in the phased-array probe is pulsed sequentially, launching an ultrasonic wave that propagates through the solid specimen. Echoes from internal reflectors are recorded simultaneously by every element of the array, yielding the complete transmit-receive data matrix needed for TFM reconstruction. Each reflection within the solid produces compressional/Longitudinal (L) and Shear/Transverse (T) waves, which propagate at different velocities and along distinct trajectories. To obtain multiple acoustic signatures of the same internal defect and enhance the information content of each sample of our training data, we employ multimodal TFM (M-TFM) reconstruction. It considers both direct and indirect propagation paths as well as mode conversions between wave types. While this approach enhances image richness, it also increases complexity for manual interpretation. With recent advances in ML demonstrating human-level or superior performance in various domains, there is growing interest in applying Deep-Leaning (DL) to automate defect characterisation in NDT&E when dealing with vast amounts of data. However, applying simulation-trained DL models to experimental conditions introduces specific challenges [2].

In this work we deploy an Automated Machine-Learning (AutoML) [3] pipeline to train and tune several regression models on CIVA-simulated M-TFM data, then evaluate their performance and robustness on independent experimental data to analyse environmental and operational uncertainties.

The remainder of this paper is organised as follows. First, we describe the experimental setup and the datasets generated from it. Next, we present the DL pipeline and report the corresponding results. Finally, we draw our conclusions and outline directions for future work.

Experimental Setup

The experimental specimen is a ferritic-steel mock-up that replicates the geometry of a butt-weld joint as used in [4]. Its back wall is deliberately complex where two planar facets meet at the weld root and slope in opposite directions (view Fig. 1). Four artificial notches have been machined into the chamfer to represent surface-breaking cracks. Each notch is 0.2 mm wide and 20 mm long. Their heights alternate between 3 mm and 10 mm, with the first two oriented vertically and the last two tilted 14° from the vertical axis. The inspection is carried out by coupling a 64-element, 5 MHz linear array (0.6 mm pitch) through a 15 mm-high Rexolite wedge set at

DOI 10.5162/Ultrasonic2025/E4-a3

37° in contact to the specimen, which refracts the beam to roughly 55° inside the steel. The material is treated as isotropic, with longitudinal and shear-wave velocities of 5 920 m/s and 3 230 m/s, respectively. This arrangement provides a realistic weld-inspection scenario in which direct, mode-converted, and back-wall-reflected paths all interact with the sloping rear surface and the surface-breaking cracks. The Panthere acquisition system from Eddyfi Technologies has been used to excite the piezoelectric elements of the probe using a sampling frequency of 50 MHz.

Previous Work

Four experimental acquisitions were performed, each centered on a different notch, with the wedge's back side positioned at X_0 = 30 mm from the specimen edge, as illustrated in Fig. 1. To model operational uncertainties representative of real inspections, we reconstructed altered TFM images from the same FMC data using nine combinations of T-wave velocity and back-wall slope within the specimen's tolerance range. The simulated dataset is composed of parameter variations close to the real mock-up settings, resulting in 6000 training, 550 validation, and 900 test samples. Inputs to the machine learning models are nine M-TFM reconstructions (TT, TTT, TTL, TLT, TLL, TTTT, TLLT, TTLT, MAX – pixel-wise maximum of the first eight modes). The training labels used for the supervised regression task consist of four parameters (Flaw height, Flaw tilt, Backwall slope, Celerity T-waves) directly derived from the specimen's characteristics sensed by the FMC acquisitions (FMC parameters), and two parameters (Backwall slope, Celerity T-waves) associated with the reconstruction settings of the TFM imaging algorithm (TFM parameters). In total, we obtain six different parameters to predict as shown in Table 1. The FMC parameters correspond to physical alterations of the model specimen chosen to be similar to the experimental mock-up, whereas the TFM parameters are obtained by perturbing the reconstruction settings to model plausible environmental uncertainties.

Recent Work

This work aims to extend the previous experimental campaign by collecting newly acquired experimental data to study the effects of both environmental and operational conditions. Four lateral scans of the probe along the butt-weld were performed, each acquiring one FMC every 2 mm. Between each scan, the probe was incrementally moved 2 mm closer to the weld to analyze the effect of positional uncertainty relative to the butt-weld. In the simulated training dataset, the probe was positioned at a fixed distance (X_0=30 mm). By progressively decreasing the distance between the probe and the weld in the experimental acquisitions to imitate operational uncertainties, we aim to assess the impact of these variations on TFM reconstructions and prediction performance. Furthermore, we repeated the acquisitions while introducing 0°, 1°, 2°, and 3° of skew to the probe to evaluate the model's regression performance under angular positioning errors, as may occur in real inspection scenarios. In all cases, we preserved the same input/output data structure as previously.

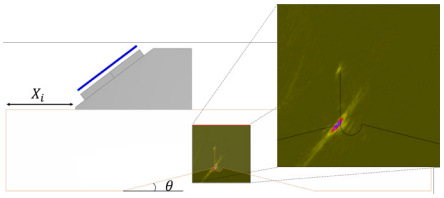

Fig. 1: Plane view of the mock-up showing the ultrasonic probe mounted on a wedge in contact with the specimen. The reference position is set at X_0 = 30 mm, which places the probe's front edge 20 mm from the weld root. The backwall slope is at an angle θ=14°. The TT-mode TFM reconstruction is displayed, revealing the 10 mm vertical defect located at the weld root.

Deep-Learning Pipeline

To streamline and automate the training of machine learning models across different mock-up datasets, we developed a pipeline that combines state-of-the-art (SOTA) model backbones with an AutoML-based backend to optimize their fully connected layers. Specifically, the deep learning pipeline handles model selection, hyperparameter tuning, and full training of the most promising architectures. In our case, four model backbones (i.e., ResNet-50, DeiT, EfficientNet-B3 and VGG-16) implemented via the PyTorch Image Models (timm) library [5] are adapted for regression by replacing their classification heads with a fully connected block, whose depth (1–3 layers) and width (128, 256, 512, or 1024 neurons per layer) are determined through AutoML. We use Ray Tune [6] to perform randomized hyperparameter search, varying additional factors such as the learning rate (10^{-4} to 10^{-2}, log-uniform distribution), batch size (32 or 64), dropout rate (0 to 0.5, uniform distribution), and whether batch normalization is activated. Each trial runs for a few epochs on the simulated training set, with validation mean-squared error as the objective loss type in order to determine the promising model variants. Once the training procedure is ended, the top-K models are chosen and retrained from scratch for a higher number of epochs, using

DOI 10.5162/Ultrasonic2025/E4-a3

Tab. 1: Parameters Used to Simulate and Reconstruct Data for Training

Parameters	Range of Variation acquisition parameters (FMC)	Range of Variation reconstruction parameters (TFM)
Flaw height [mm]	[2.0, 12.0]	NA
Flaw tilt [deg]	[-20, 0]	NA
Backwall slope [deg]	[10, 18]	[10, 18]
Celerity T-waves [m/s]	[3030, 3380]	[3030, 3380]

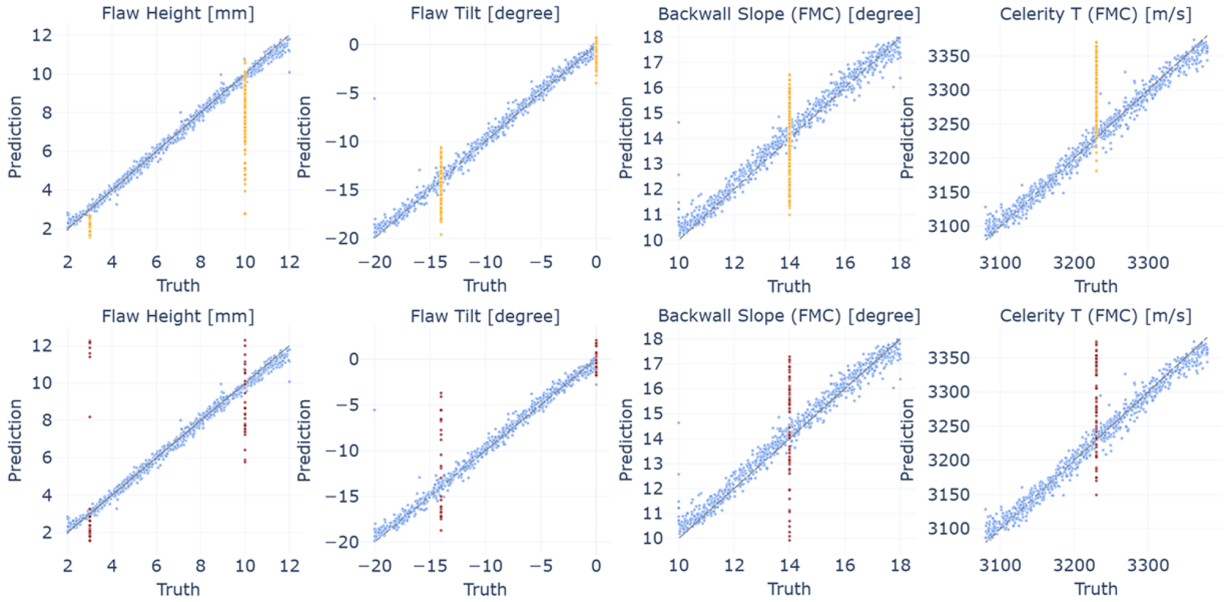

Fig. 2: True vs predicted plots comparing model outputs to ground truth values for flaw height, flaw tilt, back-wall slope, and T-wave celerity. Blue points correspond to simulated data, yellow points (top row) to experimental data with environmental uncertainties, and red points (bottom row) to experimental data with operational uncertainties. The grey dashed diagonal represents perfect prediction ($y = x$), with deviations indicating prediction errors.

early stopping with a patience threshold of 20% to prevent overfitting. Once trained, the models are frozen and evaluated on the experimental dataset. Model performance is reported using the mean absolute error (MAE), which represents the average absolute difference between predicted and true values: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ and the coefficient of determination (R^2), which reflects how well the predictions approximate the true values: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$. All training procedures use the Adam optimizer.

Model customization

To maintain a general-purpose pipeline, we included only a minimal set of preprocessing layers. First, a normalization layer independently adjusts each TFM image, giving the model resilience to gain variations introduced by the operator or acquisition system. Second, a global max-pooling layer is appended to the end of the backbone, enabling the model to handle varying image resolutions seamlessly. These modifications are common in the machine learning community and preserve the generality of our pipeline, making it applicable to other related inspection problems.

Results analysis

In this work, we compare two data acquisition scenarios in Fig. 2. The first involves previously acquired data with a single FMC acquisition per flaw, where only TFM reconstruction uncertainties are considered, specifically variations in T-wave celerity and backwall slope, representing environmental conditions. The second scenario, based on newly acquired data, introduces operational uncertainties such as probe skew and varying the probe's distance from the butt-weld. Each flaw is scanned at four probe distances (0, 2, 4, and 6 mm) as shown in Fig. 1 and four skew angles (0°, 1°, 2°, and 3°). For each of the 16

DOI 10.5162/Ultrasonic2025/E4-a3

Tab. 2: MAE and R^2 metrics for flaw height and tilt versus probe skew and position increment

	Flaw tilt								**Flaw height**							
	MAE [deg]				R^2 [a.u.]				MAE [mm]				R^2 [a.u.]			
Skew variations [deg]	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Increment 0 mm	3.92	4.06	4.65	4.97	0.49	0.46	0.25	0.21	2.16	2.14	3.80	3.36	0.54	0.49	−0.88	−0.89
Increment 2 mm	1.18	1.69	3.01	3.13	0.96	0.92	0.67	0.65	1.46	1.28	3.23	3.26	0.81	0.84	−0.68	−0.87
Increment 4 mm	1.77	1.83	1.72	1.52	0.91	0.91	0.91	0.92	1.40	1.00	1.07	3.58	0.81	0.87	0.86	−0.92
Increment 6 mm	2.27	2.37	2.13	2.32	0.88	0.87	0.89	0.85	1.29	1.14	1.82	2.90	0.80	0.84	0.41	−0.68

distance–angle scan configurations, we performed an independent FMC acquisition on the center of each of the four flaws, yielding $16 \times 4 = 64$ samples (red points in Fig. 2). This setup allows us to evaluate the impact of variable operational conditions on regression performance.

We trained 100 models per architecture on simulated data and selected the top performer for further evaluation. During the hyperparameter search phase, we used RayTune [6] with randomized sampling to explore configurations across all four architectures. In our experiments, VGG-16 achieved the best overall predictive performance, closely followed by the others. Fig. 2 illustrates its results: The simulated set, generated from randomized variations, is visualized as blue points aligning closely with the theoretical diagonal, demonstrating high prediction accuracy. For the previously acquired experimental data containing environmental uncertainties, TFM reconstructions were generated for each of the four flaws by combining nine TFM variations for both the backwall slope and the T-wave celerity, using the same FMC acquisition, resulting in 324 $(4 \times 9 \times 9)$ prediction samples. The gap in predictive performance on experimental data stems both from the lack of varying increment and skew effects in the training set and from discrepancies between simulated and real ultrasonic images, and it is supposed that these discrepancies migth be properly addressed by employing fully numerical (i.e., finite element method) solvers. In Fig. 2, we further evaluated the same model on a second experimental dataset, which included variations in probe distance from the butt-weld and different skew angles. In contrast, varying the probe's increment for the experimental acquisitions produced only marginal changes in performance. In Table 2 we can see the performance drop depending on the probe increment and skew variations, where the latter seems to have a higher impact on the prediction results.

Conclusion

Multiple directions emerge from this study. First, we showed that a machine learning model trained solely on simulated data can achieve reasonable performance on experimental data for regression-based inversion problems using readily available SOTA backbones without extensive model optimization. Second, our AutoML pipeline successfully trained multiple state-of-the-art models tailored to the task, with performance on both simulated and experimental data comparable to previous work. Despite lower accuracy on experimental data due to domain differences and real-world uncertainties, the pipeline is easy to implement and was evaluated under environmental and operational conditions representative of on-site inspections. Future work will focus on developing DL strategies to leverage a subset of acquisitions, integrating simulations that account for operational conditions, and improving the model's experimental performance through the established pipeline.

References

[1] C. Holmes, B. W. Drinkwater, and P. D. Wilcox. "Post-processing of the full matrix of ultrasonic transmit–receive array data for non-destructive evaluation". In: *NDT & E International* (2005).

[2] G. E. Granados et al. "Generative domain-adapted adversarial auto-encoder model for enhanced ultrasonic imaging applications". In: *NDT & E International* (2024).

[3] M. Baratchi et al. "Automated machine learning: past, present and future". In: *Artificial Intelligence Review* (2024).

[4] R. Miorelli et al. "Use of deep learning and data augmentation by physics-based modelling for crack characterisation from multimodal ultrasonic TFM images". In: *Nondestructive Testing and Evaluation* (2024).

[5] R. Wightman. *PyTorch Image Models (timm)*. 2019. URL: https://github.com/huggingface/pytorch-image-models.

[6] R. Liaw et al. "Tune: A Research Platform for Distributed Model Selection and Training". In: *arXiv preprint arXiv:1807.05118* (2018).