applied sciences

MDPI

Article

Determination of Groove Filling Levels of Pressed Pipe-Fitting Connections Using Phased Array Ultrasound Evaluated by a CNN

Kevin Jacob [1,2], Benjamin Straß [2,*], Nico Brosta [2] and Jaqueline Presti-Senni [2]

[1] Chair of Cognitive Sensor Systems, Saarland University, 66123 Saarbrücken, Germany; kevin.jacob@izfp-extern.fraunhofer.de
[2] Fraunhofer Institute for Nondestructive Testing IZFP, 66123 Saarbrücken, Germany; nico.brosta@izfp.fraunhofer.de (N.B.); jaqueline.presti-senni@izfp.fraunhofer.de (J.P.-S.)
* Correspondence: benjamin.strass@izfp.fraunhofer.de

Abstract

In this paper, a method for determining the filling level of grooves (1 mm (W) × 0.25 mm (H)) in pressed titanium pipe-fitting joints is presented. The joints are inspected in a water bath using a 20 MHz phased array ultrasound, and the acquired raw B-scans are evaluated by a convolutional neural network that performs per-groove regression. Reference filling levels are obtained destructively from micrographs. Compared to X-ray computed tomography and destructive sectioning, the proposed approach overcomes the low material contrast between pipe and fitting, avoids long scan times, and enables a nondestructive, potentially inline-capable quantitative assessment of sub-millimeter grooves. A manual high-frequency ultrasound evaluation with a single probe and conceivable rule-based time-of-flight pipelines with hand-crafted echo picking and thresholds both show only moderate agreement with CT references and require substantial feature engineering for multiple echoes. In contrast, the PAUT-CNN method exploits the full raw B-scan without explicit feature design and achieves a root mean square error of about 7% of the groove filling levels on a held-out test set, corresponding to an absolute error on the order of a few tens of micrometers in groove height. This demonstrates that high-frequency phased array ultrasound combined with data-driven evaluation can quantitatively assess the filling of sub-millimeter grooves in aerospace-relevant press-fit connections.

Keywords: ultrasound; phased array; convolutional neural network; pressed pipe fittings; groove filling level; NDT; NDE

Academic Editor: Andrea Carpinteri

Received: 30 December 2025
Revised: 20 February 2026
Accepted: 24 February 2026
Published: 26 February 2026

1. Introduction

Pressed pipe-fitting connections are critical components in aerospace and other high-performance applications. In order to ensure structural integrity, sealing capability, and process safety while maintaining low weight, nondestructive evaluation (NDE) is inevitable. Faulty connections can lead to leakage, structural failure, or costly rework. The groove filling level directly correlates with connection quality and service life. Conventional quality control is often carried out destructively by micrographs, which are time-consuming and costly. Furthermore, it is impossible to achieve 100 percent inspection using these methods. Process variability, such as that due to pressing force, material tolerances and tool wear, requires an inline-capable NDE solution.

The subject of this work is the examination of pressed pipe-fitting connections. We present the main results of the Project "Hauptarbeitspaket 2–NDT-Verfahren (Einrollen)" [1]. The fittings have six annular grooves on the inside that run around the circumference with a cross-sectional area of 1 mm × 0.25 mm. A sketch of the cross-section of such a pipe-fitting connection is shown in Figure 1. The filling level of the grooves was defined as shown in Figure 2. Considering a single groove, x is defined as the difference between the highest and lowest point on the pipe and y as the difference between the highest and the lowest point of the fitting. The filling level of the groove is then given as $F = x/y \cdot 100\%$.

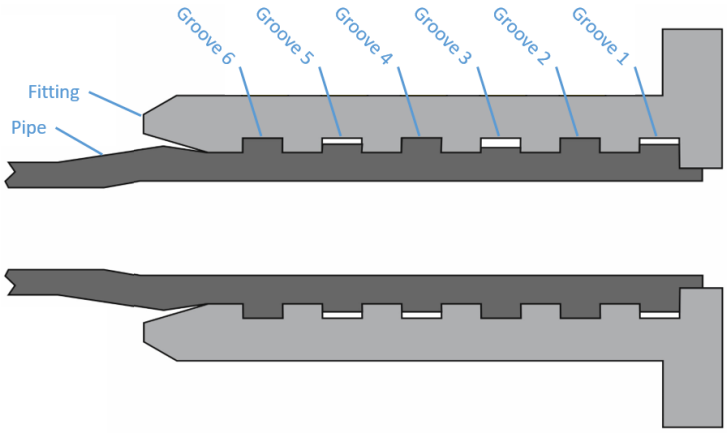


Figure 1. Schematic cross-section of a pipe-fitting connection with grooves.

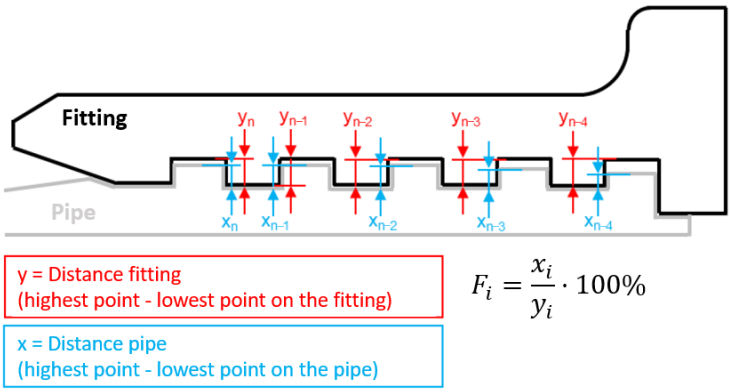
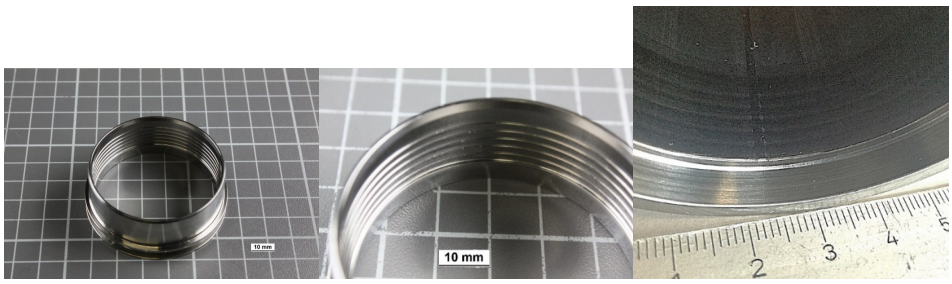


Figure 2. Definition of the groove filling levels.

The pipes consist of pure titanium, and the fittings consist of a titanium alloy. Photographs of a fitting are shown in Figure 3a,b, and in Figure 3c, a photograph of a pressed pipe-fitting connection taken from the inner side of the pipe is shown. The imprints of the tool for pressing the inner pipe into the grooves of the fitting are clearly visible.



(**a**) (**b**) (**c**)

Figure 3. Photos of the specimens: (**a**,**b**) photos of the fittings with annular grooves on the inside. (**c**) Photo of a pressed pipe-fitting connection.

For the joining of the pipe-fitting connections, the fittings are placed on the pipes, which are then pressed into the fittings resp. The grooves of the fittings from the inside. The task is to nondestructively determine the filling level of those grooves.

In aerospace fluid-distribution systems, pressed titanium pipe-fitting connections of the type considered here are used in fuel, hydraulic, and bleed-air lines where tightness, low mass, and high reliability are mandatory. In current industrial practice, the quality of such joints is typically qualified by destructive sectioning of a limited number of specimens and, where feasible, by high-resolution X-ray CT on representative samples, while routine production relies mainly on process control and leak testing. To our knowledge, there is no established nondestructive technique that provides quantitative groove-by-groove filling levels for these titanium press-fit joints under production conditions. The present work, therefore, addresses an industrially relevant gap by developing a PAUT plus CNN route that yields per-groove filling values that can, in principle, be integrated into automated test stations for aerospace pipe manufacturing.

2. State of Art and Related Work

The state-of-the-art solutions for the given inspection task include X-ray computed tomography (CT), which allows for a complete 3D visualization with a high resolution. X-ray imaging measurements combine X-ray exposure with computational reconstruction to produce images or sequences of the object studied [2]. In contrast to radiography 2D outputs, X-ray CT provides 3D models by taking many radiographs as the specimen rotates between the source and the detector, followed by model reconstruction and visualization [3]. This technique delivers a high spatial resolution, making X-ray methods particularly effective for nondestructively identifying most irregularities in materials and welds. The process can be time-consuming and, due to the required rotation, may be restricted by geometry, depending on material and thickness. Therefore, radiography and CT are often used as reference methods [3,4]. X-ray techniques can detect very small defects within the specimen's volume and can also be used to characterize different material or joint properties beyond defect detection [5]. The disadvantages in this case are, on the one hand, the similar X-ray attenuation of Ti and Ti-alloy, which are often used for such joints in aerospace applications and results in low material contrast. Further, it would require long measurement times in the range of hours to days by means of μCT to achieve sufficient voxel resolution in the order of 10 μm, which makes an inline integration impossible. For these reasons X-ray CT examinations were used solely as reference methods to visualize the joint quality of selected test specimens and thus ultimately to assess the detection limits of the other NDT methods within the described project. A possible remedy for these shortcomings is provided by ultrasound procedures, although the quantitative filling level determination in sub-millimeter grooves remains barely documented.

With respect to the achievable resolution and detection sensitivity, conventional ultrasound techniques often cannot detect sufficiently small irregularities required for efficient and safe lightweight design structures. To address this, specialized high-frequency ultrasound methods (HF-US) have been developed, operating in the 10 to 200 MHz range. The higher ultrasound frequency yields an improved resolution and sensitivity, which can be achieved by performing measurements in a water-filled immersion tank, taking advantage of the velocity difference between water and air [6]. As a result, irregularities with a size up to 0.2 mm and even kissing bonds in weld seams, which are very difficult to detect with other NDT methods, can be reliably detected [5]. In contrast to the measurements using a single transducer as a probe, phased array ultrasonic testing (PAUT) uses arrays of probes. An array consists of several elementary transducers that can be controlled independently of each other. They generate spherical (matrix array) or cylindrical (linear

array) waves that overlap in the material and form a wavefront through their interference. By exciting the elements at different times, the shape of the interference can be controlled and thus swiveled or focused [5,7]. Nevertheless, HF-US and PAUT are not applicable as inline NDT-methods for a lot of applications due to corrosive effects. In the case of the application described in this paper, no adverse effects due to corrosive effects are to be expected, which is why the groove fill level was determined using PAUT.

For titanium alloys, ultrasonic propagation is characterized by comparatively high attenuation and microstructural scattering at high frequencies, as well as a relatively small acoustic impedance contrast between different titanium grades. At 20 MHz, this limits the inspection depth but remains sufficient for the wall thickness of the present pipes while, at the same time, making purely amplitude-based defect detection challenging. Phased array immersion testing was, therefore, chosen over single-element probes or contact setups because electronic focusing and steering enable the concentration of energy in the groove region and the acquisition of 2D B-scans in a single rotation. These B-scans contain multiple reflection paths and later echoes that are particularly valuable for distinguishing different groove filling levels in low-contrast titanium joints.

The ever-increasing use of AI in NDE also contributes to new possibilities and more robust evaluations of NDE sensor data. In recent years, there has been an increase in research focusing on the AI-based evaluation of ultrasound data for various inspection situations. Several studies support the feasibility of operating on minimally processed PAUT data. Siljama et al. show that CNNs can ingest multi-channel PAUT B-scans without the synthetic aperture focusing technique (SAFT) or the total focusing method (TFM) to detect weld flaws, leveraging augmentation to scale training [8]. Virkkunen et al. similarly train deep CNNs on immersion PAUT B-scans around 1.8–2 MHz, using extensive virtual flaw augmentation while avoiding reconstruction [9]. Pushing toward even rawer inputs, Jia and Rakhmatov classify crack attributes directly from 2D raw channel frames, highlighting the value of phase-preserved measurements over beamformed images [10]. These works collectively demonstrate robust learning on raw or near-raw PAUT data, but they focus on detection or attribute classification. Continuous regression from array ultrasonics has been demonstrated most convincingly on reconstructed images. Pyle et al. use CNNs on plane-wave images at 5 MHz to regress crack length and angle, outperforming conventional sizing with hybrid simulated and experimental datasets [11]. This establishes the practicality of quantitative targets in ultrasonic ML. Complementary evidence for quantitative characterization from coherent raw representations comes from Bai et al., who compare ML and Bayesian inversion on scattering matrices derived from FMC scans, arguing that physics-aware, phase-coherent domains can support parameter estimation with uncertainty [12]. Together, these precedents suggest that regression is feasible and that raw or coherent representations are advantageous. Although not targeting regression from raw B-scans, this line of work underscores the acoustic and calibration demands of high-frequency arrays in Ti that are directly relevant at approximately 20 MHz. As a geometry analog, Shi et al. classify inner-wall circumferential slots using raw A-scans at around 2 MHz, demonstrating the practicality of circumferential scanning with a probe aligned to the pipe, but without PAUT, without regression, and at a far lower frequency [13]. For broader context, two additional references illustrate the surrounding landscape without directly advancing the target. Latete et al. explore CNNs for PAUT defect location, identification, and sizing, contributing to the general trend toward ML-driven characterization in array ultrasonics, but without raw RF B-scan inputs, titanium-specific setups, or per-feature continuous regression [14]. Naddaf-Sh et al. benchmark transformer and YOLO detectors on industrial PAUT B-scan images of pipeline welds, strengthening image-level detection baselines on real data but not addressing raw RF inputs, titanium

materials, or continuous regression of quantitative targets [15]. In parallel, several studies have investigated the ultrasonic testing of titanium alloys and quantitative characterization of small geometric features such as shallow surface-breaking notches or narrow grooves in metallic components [16,17]. These works demonstrate that sub-millimeter defect sizing in titanium is feasible in principle, but they focus on different joint geometries and do not provide continuous per-groove filling levels in press-fit connections. The present study complements this literature by targeting closed sub-millimeter grooves in titanium pipe-fitting joints and by combining high-frequency PAUT with CNN-based regression of the groove filling. In summary, the closest building blocks to our goal are raw multi-channel or channel-frame ingestion without reconstruction for detection or classification [8–10], continuous regression for ultrasonic characterization but on reconstructed images [11], coherent raw-domain parameter estimation with uncertainty [12], and circumferential inner-wall slot scanning as a geometry analog at a low frequency [13]. To our knowledge, within this set, there is no prior demonstration of continuous 0–100 percent per-groove fill regression from raw PAUT B-scans in titanium press-fit connections. The present work addresses this gap by combining phase-preserved raw B-scan inputs, high-frequency Ti acquisition, and per-groove CNN regression grounded in destructive metrology while situating results against image-level detection baselines and general PAUT-CNN sizing efforts for context [14,15].

Compared to conventional inspection routes, the proposed PAUT-CNN approach offers several advantages for the present titanium press-fit joints. X-ray computed tomography is limited by low contrast between pipe and fitting, requires long scan times, and is therefore unsuitable for inline use in this context. Destructive micrographs provide accurate groove filling levels but are slow, costly, and cannot be applied to every joint. Manual high-frequency ultrasound evaluation with a single probe, as illustrated in Table 1, yields only moderate agreement with CT references and does not scale well because it relies on hand-picked transit times. A more automated time-of-flight evaluation based on hand-crafted features and multi-echo rules would be possible in principle, but it would require complex and inspection-specific signal processing pipelines that are difficult to tune and to maintain under varying noise and coupling conditions. In contrast, the PAUT-CNN method exploits the full raw B-scan, including later echoes, learns the relevant patterns directly from data without explicit feature engineering, and achieves a test RMSE of about 7% across all grooves while remaining fully nondestructive.

Table 1. Results of manual evaluation of single-sensor measurements from the inside of a pipe-fitting connection. Red indicates results that were identified as incorrect in the manual evaluation.

Groove	Probe Position Groove	Probe Position Adjacent Bump	Calculated Distance Between Entry Echo and Groove Echo (in mm)	Calculated Distance Between Entry Echo and Bump Echo (in mm)	Difference = Determined Groove Filling (in mm)	Reference Value from CT Data (in mm)
6	2	1	0.70	0.65	0.05	0.07
6	2	3	0.70	0.645	0.055	0.07
3	6	5	0.72	0.626	0.094	0.113
3	6	7	0.72	2.287	−1.567	0.113

The application of convolutional neural networks to raw sensor data follows the broader paradigm of deep learning, where hierarchical representations are learned directly from input data through end-to-end optimization [18]. This approach has demonstrated superior performance across diverse domains by eliminating manual feature engineering in favor of data-driven feature extraction through multiple layers of nonlinear transformations. For ultrasonic NDE, this paradigm shift enables the network to discover complex acoustic

patterns (multi-echo interference, phase relationships, geometric signatures) that would be difficult to encode through conventional signal processing rules, particularly for sub-millimeter features in titanium joints where traditional time-of-flight analysis showed limited accuracy.

3. Methodology

In the first step, a pipe-fitting connection was tested from the outside of the pipe using phased array ultrasound in order to obtain an initial indication of the detectability of the grooves. The test setup for performing the measurements on the connections in question is shown in Figure 4. The test was carried out using the immersion technique with a 20 MHz phased array probe with a linear scan.

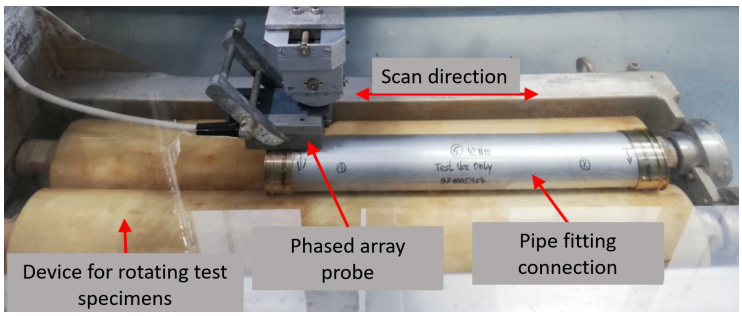


Figure 4. Test setup for phased array testing using HF-US from the outside of the pipe.

Before the groove filling levels were determined destructively, reference measurements by means of X-ray computed tomography (CT) were obtained. The resulting voxel edge length (38 µm) of the reconstructed volume image was unfortunately too low to get groove filling levels with satisfying precision. Quantitatively, the groove height of 0.25 mm corresponds to only about six to seven voxels in the reconstructed CT volume, so partial-volume effects and small segmentation inaccuracies translate into large relative errors in the estimated filling level. Together with the almost identical X-ray attenuation of the titanium pipe and the titanium-alloy fitting, this prevents reliable quantitative per-groove filling values from CT, so in the present study, CT is used only qualitatively to verify that the intended variation of groove filling across specimens has been achieved and to visualize the overall joint morphology. Furthermore, the pipe and the fitting cannot be distinguished in CT scans due to identical X-ray attenuation, which made it even more difficult to determine the groove filling levels. CT sectional views for pipe-fitting connections manufactured with low, medium and high force, respectively, are shown in Figure 5a–c. The targeted variance in the filling levels of the grooves of the differently pressed pipe-fitting connections could thus be verified, as can be seen clearly.

The ultrasound A, B and C scans were used for evaluation. First, a measuring aperture was placed over the rear wall echo of the inner component (pipe), with the signal level scaled to 80% (see red frame in Figure 6). This clearly shows the areas where the sound signal reaches or does not reach the inner pipe. The ultrasound B scan also shows that the rear wall of the fitting is partially invisible, whereas, in these cases, the rear wall of the pipe is clearly visible.

This already indicates the quality of the connection since, in the event of an air gap in the grooves, the sound transmission is interrupted, and therefore, no rear wall echo of the pipe can be detected within the area outlined in red. The red dotted line on the right-hand side of Figure 7 illustrates the measurement position of the tested connection, which is shown in the amplitude image on the left.

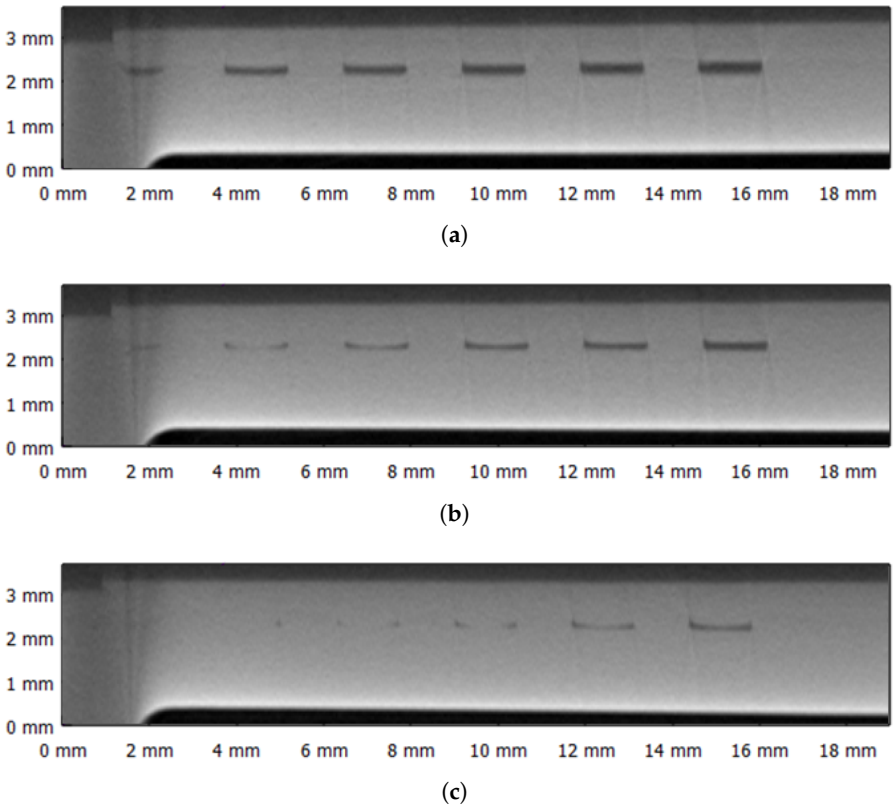


Figure 5. CT images of pipe-fitting connections: (**a**) Pressed with low force. (**b**) Pressed with medium force. (**c**) Pressed with high force.

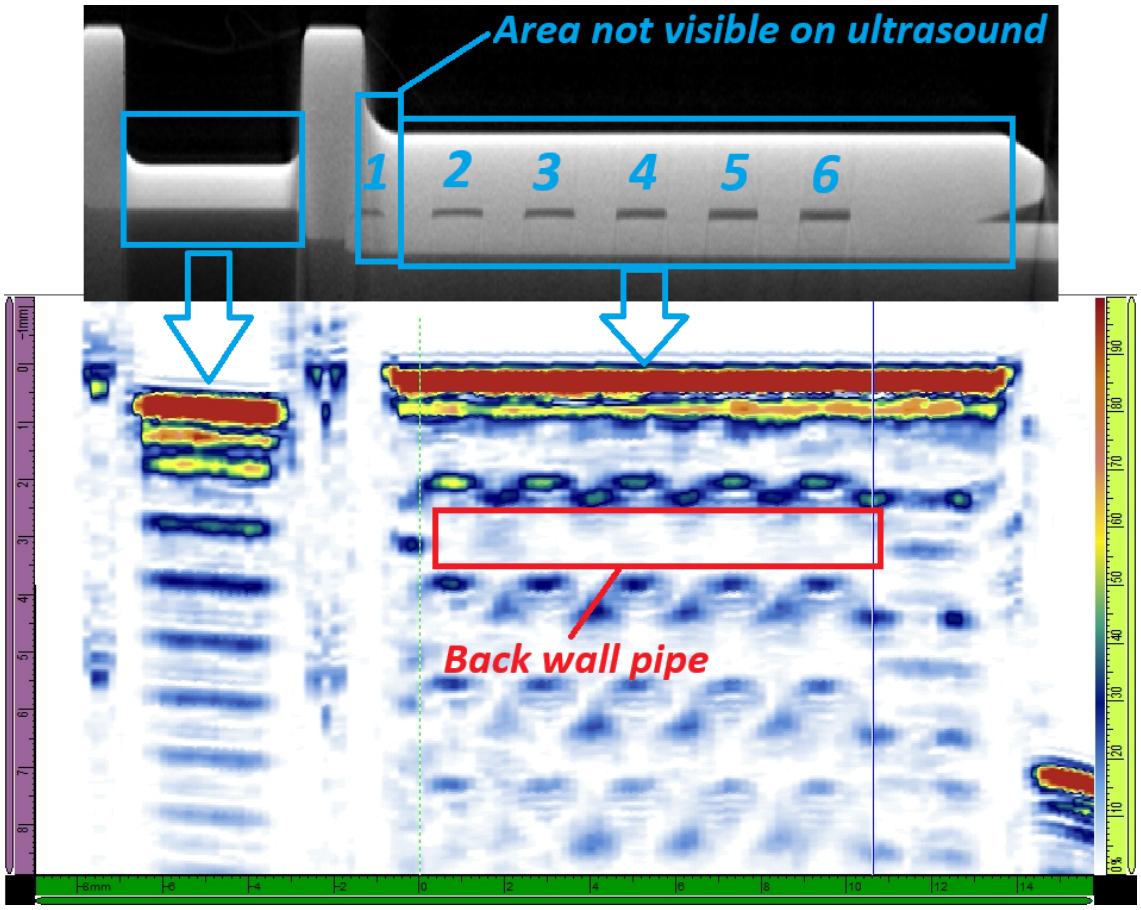


Figure 6. Identification of relevant areas in the ultrasound image.

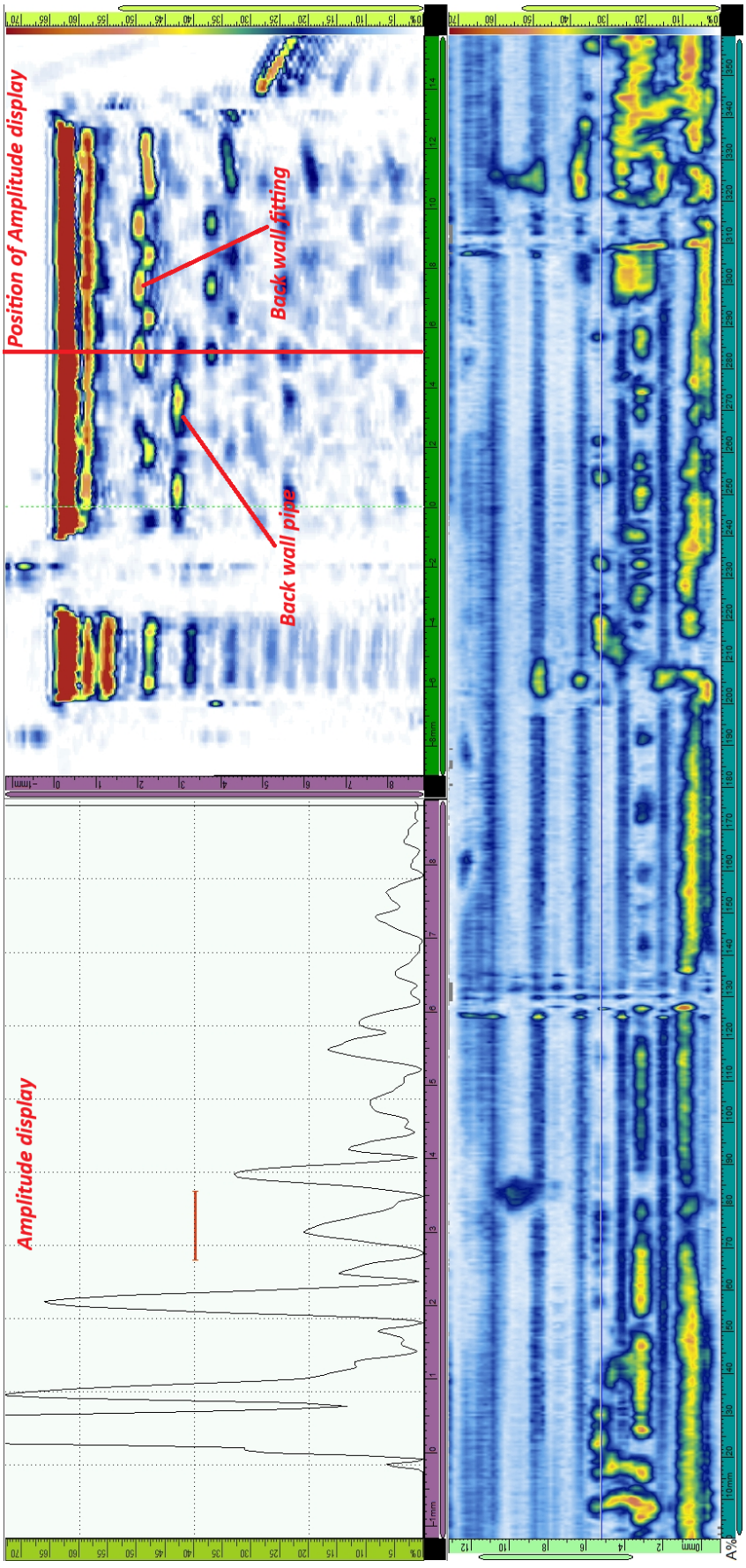


Figure 7. Procedure for evaluating the measurements.

It should be noted that the first groove (directly at the transition to the larger diameter of the fitting (on the left in each figure)) cannot be detected here due to the geometric boundary conditions. Since the first groove could not be detected from the outside of the pipe in the previous phased array measurements due to geometric effects, investigations were carried out to determine possible optimizations of the sensor technology. To this end, after a pipe was cut open, high-resolution measurements were carried out using a single sensor from the inside of the pipe; see Figure 8. Clear echoes of the first groove were also detected in the measured signals, as can be seen in Figure 9.

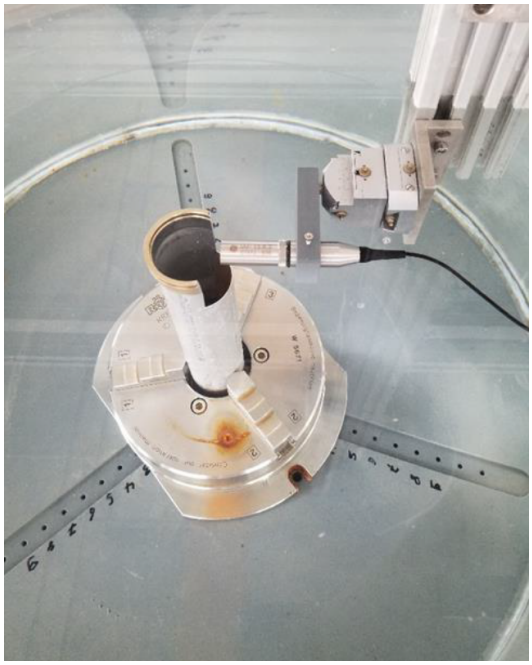


Figure 8. Measurement setup for runtime measurements with a single sensor from the inside of the pipe.

A manual evaluation of these measurements, shown in Figures 9 and 10, in which the transit times of the echoes from the grooves were compared with the transit times of the echoes from the adjacent bumps, revealed moderate agreement with the reference values derived from the CT measurements. Only in the case of transmission through the transition between the pipe and the fitting did this type of evaluation reveal discrepancies, as can be seen in Figure 10 at probe position 7 ("PK 7") and listed in Table 1.

Since the results of the manually evaluated single-sensor measurements showed a moderate agreement with the reference values, it was obvious not only to consider the first echo that falls back but also to take into account later repeated echoes to improve accuracy. The otherwise complex feature engineering required to consider multiple echoes motivated the use of a convolutional neural network for the task of determining the filling level from the ultrasonic measurement data. Since the training of a neural network requires, in general, large amounts of suitable data to achieve satisfactory accuracy, 25 pipe-fitting connections, which were manufactured with different pressing forces in order to ensure an even distribution of filling levels, were examined by means of phased-array ultrasonic measurements in a water bath. Measurements were taken at six equidistant angles around the circumference, such that, in total, 150 B-Scans were acquired. Afterwards, the pipe-fitting connections were examined destructively to determine the filling levels of the grooves at those angles. The acquired B-Scans, together with the destructively determined groove filling levels, were used to train a convolutional neural network.

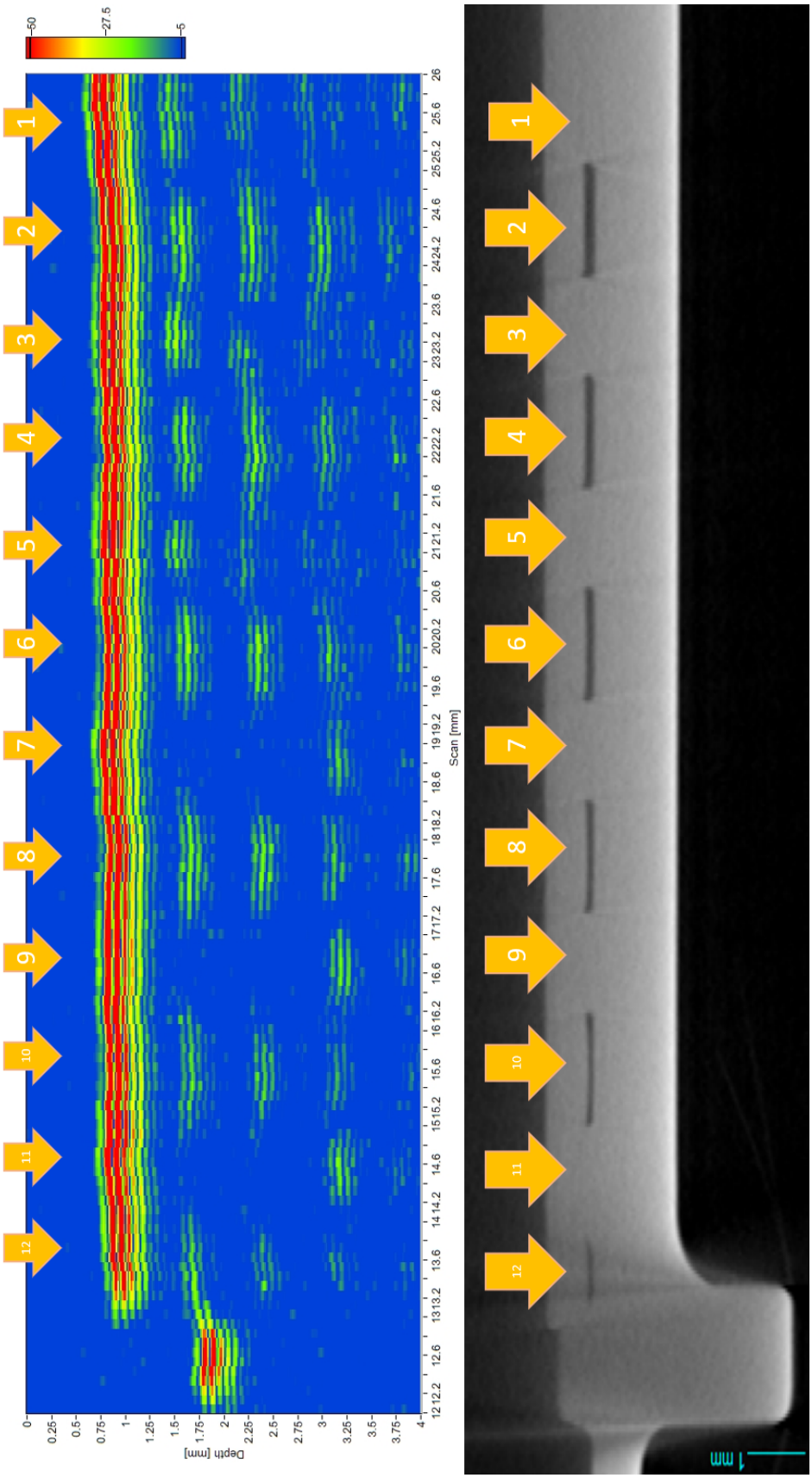


Figure 9. Comparison of HF-US-B image (**left**) and CT cross-section (**right**) of a pipe-fitting connection with several probe positions (numbered 1 to 12) marked (yellow arrows).

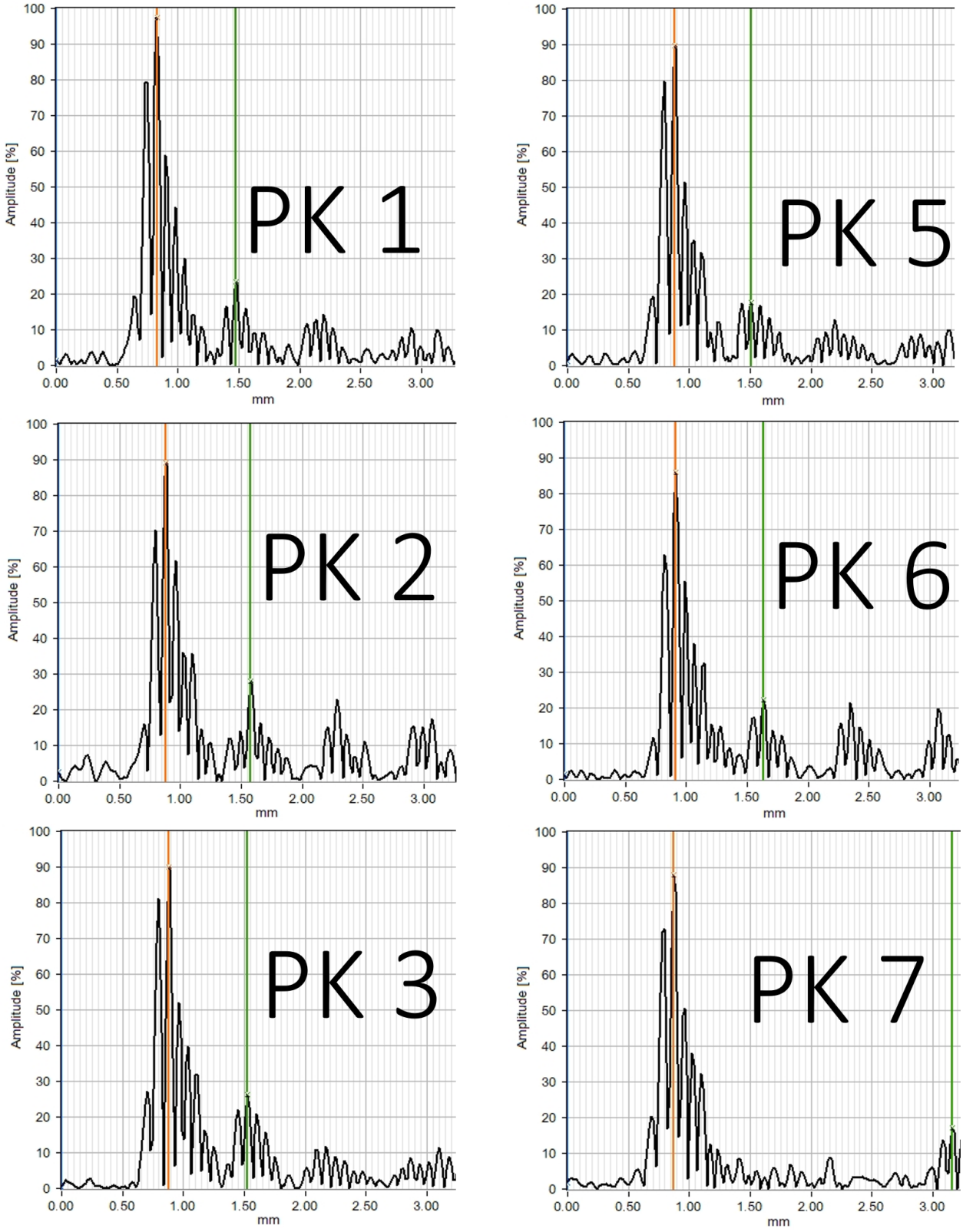


Figure 10. Manually evaluated transit time measurements for probe positions 1, 2 and 3 (**left**) and 5, 6 and 7 (**right**) of the single-sensor measurements. The orange line indicates the entry echo, while the green line indicates the echo of the groove or the adjacent bump.

Compared to this manual time-of-flight evaluation, the subsequent PAUT-CNN approach avoids explicit feature engineering and the manual picking of individual echoes. In principle, a more automated time-of-flight pipeline could be constructed by defining

algorithmic rules for detecting several echo families, measuring their relative transit times, and combining these hand-crafted features in a regression model. However, such a rule-based design would be highly specific to the present geometry and measurement setup, would require extensive tuning to remain stable under noise and coupling variations, and would still rely on a limited set of engineered descriptors. The CNN, by contrast, learns directly from the full B-scan, including later echoes and subtle multi-echo interference patterns, which reduces the influence of ambiguous transit-time differences and improves quantitative accuracy while, at the same time, eliminating manual and rule-based feature design.

For the phased array ultrasonic measurements from the inside of the pipe, an M2M Multi2000 device was used, together with an array with 128 elements with a pitch of 0.2 mm. The sample rate was set to 100 MS/s, while the excitation frequency was 20 MHz. The measurements were taken in a water bath, with the specimen rotating axially around the array that was positioned inside the pipe. B-Scans were acquired in 60° steps around the circumference of the pipe. Examples of B-Scans taken from a pipe-fitting connection, manufactured with low force, medium force and high force, are shown in Figure 11a–c. From an inline perspective, the proposed inspection chain is compatible with typical production cycle times. Each B-scan consists of 400 × 243 samples, which corresponds to less than 0.4 MB of raw data per scan, so the data volume per joint remains small, even when six circumferential positions are inspected. With standard phased-array repetition rates and a simple rotation mechanism, the acquisition of the six B-scans (in 60° steps) for one joint can be completed within a few seconds, such that ultrasonic acquisition does not dominate the overall cycle time. The trained CNN has about 65,000 trainable parameters, and a single forward pass on one B-scan takes only a few milliseconds on a modern CPU, so the computational inference time is negligible compared to mechanical handling and data acquisition. Inline feasibility is, therefore, primarily limited by the mechanical integration of the immersion setup, rather than by the evaluation algorithm itself.

The CNN-based approach follows the paradigm established by Krizhevsky et al. for image classification [19], enabling end-to-end learning of hierarchical features directly from raw 2D ultrasonic data without manual feature engineering. This is particularly advantageous for the present task, where traditional time-of-flight analysis showed limited accuracy due to complex multi-echo patterns in sub-millimeter grooves.

The architecture of the CNN used to evaluate the acquired B scans is shown in Figure 12. Tables 2 and 3 summarize the main structural details of each layer, including kernel size, stride, activation functions, dropout rates, and the number of trainable parameters. The model consists of four identical convolutional blocks with 10 filters of size 17 by 11 and stride 2, each followed by a LeakyReLU activation with an alpha of 0.1 and a dropout layer with a rate of 0.5. A flattening layer and two dense layers with 12 and 6 units form the classification head, where the final layer uses a sigmoid activation and Gaussian weight initialization with mean 0 and standard deviation 0.2. Leaky ReLU was selected over standard ReLU to prevent dying neurons during training, particularly critical, given the limited dataset size [20]. This activation function maintains a gradient flow for negative inputs while enabling the principled weight initialization strategies that facilitate convergence in deeper networks. The last dense layer has a sigmoid activation and six output nodes corresponding to the six groove filling levels. The output values between 0 and 1 correspond to the groove filling levels from 0% to 100%.

This six-output architecture implements implicit multi-task learning, where a single shared convolutional backbone extracts features from the B-scan while the final dense layer produces independent predictions for each groove [21]. Multi-task learning can improve generalization when tasks are related, as is the case here, where all six grooves share similar

acoustic properties and geometric constraints. The shared representation forces the network to learn features relevant across all grooves, rather than overfitting to groove-specific noise, which is particularly valuable, given the 114-sample training set.

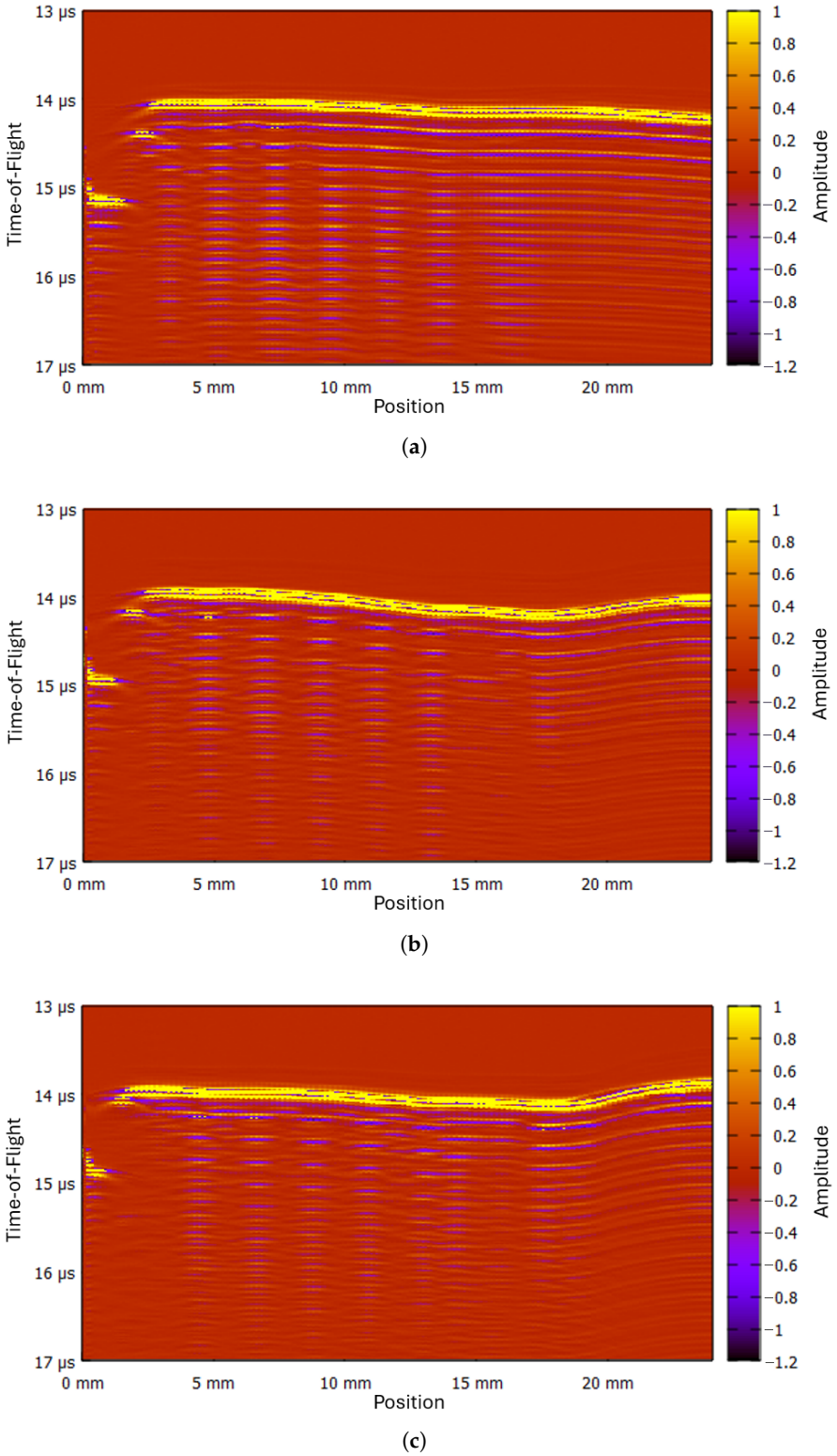


Figure 11. US B-Scans of pipe-fitting connections: (**a**) Pressed with low force. (**b**) Pressed with medium force. (**c**) Pressed with high force.

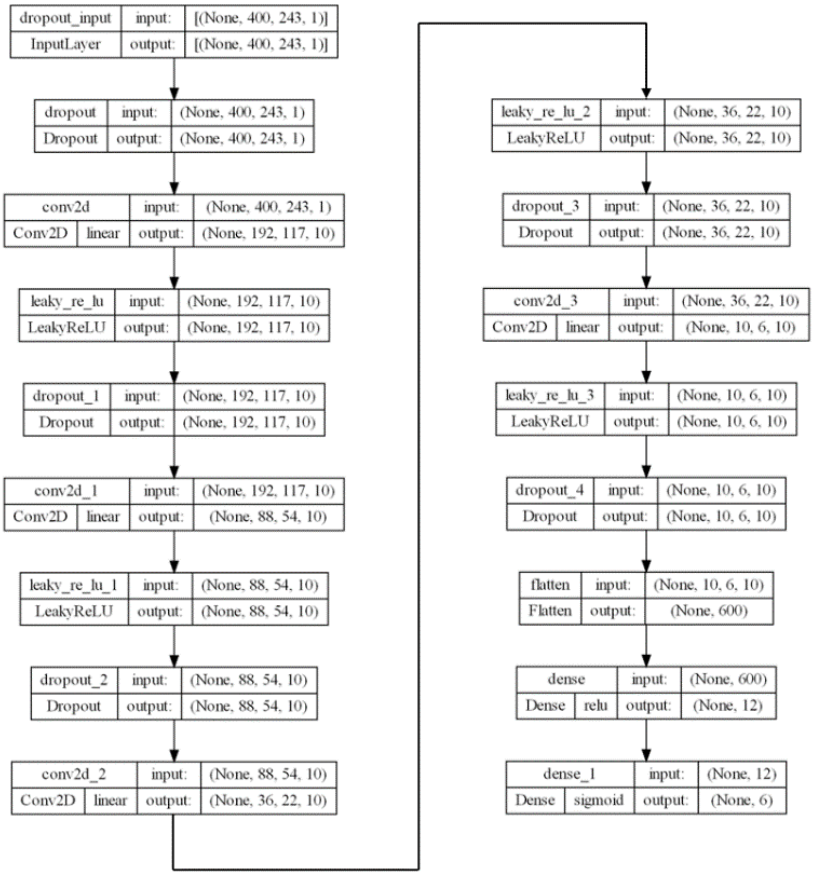


Figure 12. Architecture of the convolutional neural network employed.

Table 2. Convolutional feature extractor.

Layer	Type	Output Shape	Key Parameters	Params
Input dropout	Dropout	$(400, 243, 1)$	rate 0.5	0
Conv block 1	Conv2D, LeakyReLU, Dropout	$(192, 117, 10)$	filters 10, kernel 17×11, stride 2, alpha 0.1, dropout 0.5	1880
Conv block 2	Conv2D, LeakyReLU, Dropout	$(88, 54, 10)$	filters 10, kernel 17×11, stride 2, alpha 0.1, dropout 0.5	18,710
Conv block 3	Conv2D, LeakyReLU, Dropout	$(36, 22, 10)$	filters 10, kernel 17×11, stride 2, alpha 0.1, dropout 0.5	18,710
Conv block 4	Conv2D, LeakyReLU, Dropout	$(10, 6, 10)$	filters 10, kernel 17×11, stride 2, alpha 0.1, dropout 0.5	18,710

Table 3. Fully connected classification head.

Layer	Type	Output Shape	Key Parameters	Params
Flatten	Flatten	(600)	none	0
Dense 1	Dense	(12)	units 12, activation ReLU	7212
Dense 2	Dense	(6)	units 6, activation sigmoid, kernel init $\mathcal{N}(0, 0.2)$	78

For the training of the CNN, the total 150 B-Scans, together with the destructively determined groove filling levels, were split into 114 dates for the training data and 36 dates for the test data. The batch size was set to 1, which led to better generalization than larger batch sizes. Empirically, batch sizes of 4 and 8 reduced the stochasticity of the gradient

but consistently yielded a higher RMSE on the held-out test set, despite a slightly faster decrease in the training loss. We, therefore, treat batch size 1 as a deliberate regularization choice for this small-data regression problem, accepting the longer training time and the somewhat higher run-to-run variance in exchange for better generalization. This observation aligns with theoretical findings by Keskar et al., who demonstrated that small-batch training converges to flatter minima in the loss landscape, yielding improved generalization compared to large-batch methods that tend toward sharp minima [22]. Beyond regularization, the dropout layers enable Bayesian uncertainty estimation through Monte Carlo dropout at the inference time, where multiple forward passes with active dropout provide prediction variance [23]. This capability is particularly relevant, given the limited training set of 114 samples, as it allows the quantification of model confidence for individual predictions. Given the small dataset and regression task, the combination of batch size 1, dropout, and implicit multi-task learning was critical to achieving robust generalization. For the optimizer, Adam was chosen with a learning rate of 0.0001, while the parameter β_1 was set to 0.3 and the parameter β_2 to 0.999. The Adam optimizer [24] was selected due to its adaptive learning rate properties and computational efficiency. The non-standard β_1 value of 0.3 (default: 0.9) was chosen to reduce momentum averaging, which helped prevent overfitting, given the limited training set size of 114 samples. The mean squared error (MSE) was used as the loss function.

To quantitatively evaluate the agreement between CNN predictions and destructive reference measurements, we use the root mean square error (RMSE) of the filling levels. For a test set with N grooves and reference filling levels $F_i \in [0,1]$ and corresponding predictions $\hat{F}_i \in [0,1]$, the RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{F}_i - F_i\right)^2},$$

while the MSE, which was used as the loss function during the training, is defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{F}_i - F_i\right)^2.$$

Unless stated otherwise, we report the RMSE as a percentage, that is, the above expression multiplied by 100. In addition to the global RMSE across all grooves, we also compute separate RMSE values for each groove index to analyze potential systematic differences between groove positions.

No explicit data augmentation techniques were applied during training despite the limited dataset size of 114 B-scans. While augmentation strategies such as geometric transformations, additive noise, or intensity scaling are common for expanding small image datasets [25], preserving the physical authenticity of ultrasonic signal characteristics was prioritized. The decision avoided potential artifacts that could misrepresent echo timing, amplitude relationships, or phase information critical for acoustic interpretation. Instead, regularization through dropout and small-batch training addressed overfitting, as evidenced by the 7% RMSE on held-out test data. Future work may explore domain-specific augmentation informed by ultrasonic wave propagation physics or hybrid synthetic-experimental data generation.

In the present project, we did not employ physics-based simulation to generate synthetic B-scans because setting up and validating a full wave propagation model for the specific titanium geometry and the 20 MHz array would have exceeded the available effort. Nevertheless, high-fidelity simulation-based augmentation is a promising option for future work when aiming at larger and more diverse training sets.

The architecture does not include batch normalization layers, which are common in modern CNNs for stabilizing training and enabling higher learning rates [26]. This decision was deliberate, given the batch size of 1, for which batch normalization would compute statistics over single samples, rather than mini-batches, potentially introducing noise, rather than stabilization. With 114 training samples, the combination of dropout before each convolutional layer and small-batch gradient descent provided sufficient regularization without the statistical instability that single-sample batch normalization can introduce. Future work with larger datasets may explore batch normalization with appropriately sized mini-batches or alternative normalization strategies such as layer normalization or group normalization that are less sensitive to batch size.

The sequential convolutional architecture without residual connections was chosen for its simplicity and sufficiency, given the four-layer depth and 114-sample training set. While residual learning frameworks introduced by He et al. enable the training of much deeper networks by addressing vanishing gradients through skip connections [27], the present shallow architecture showed no evidence of gradient degradation during training. However, residual connections could be explored in future work if scaling to deeper architectures becomes necessary for more complex inspection scenarios involving multiple materials, variable groove geometries, or additional defect types beyond filling level quantification. The current architecture balances representational capacity with overfitting risk appropriate to the available data volume.

4. Results

The result of the AI training is shown in Figure 13a–h. Figure 13a shows the comparison of the destructive reference values (x-axis) with the output of the neural network (y-axis) for the training data. The comparison for the test data, i.e., the data that is not used to train the neural network but to evaluate how well the neural network generalizes to new, unseen data, is shown in Figure 13b. These generally provide a slightly poorer correlation with the reference values than the training data. For the test data, it was possible to achieve agreement between the predictions of the neural network and the reference values down to an RMSE (root mean square error) of approximately 7% of the groove filling level. In Figure 13c–h, the groove filling levels are shown separately for the different groove positions.

Table 4 summarizes RMSE, mean error, and standard deviation per groove index. The values indicate that the prediction performance is relatively uniform across grooves, with only limited variation between positions. We did not observe a clear trend of systematically higher errors near the pipe-to-fitting transition or at specific groove indices, which suggests that the network can handle the moderate geometric variations within the present configuration.

Table 4. RMSE, mean error and coefficient of determination r^2 of the CNN predictions per groove on the test set.

Groove	RMSE [%]	Mean Error [%]	Standard Deviation [%]	r^2
1	6.54	1.73	6.31	0.8996
2	6.21	0.96	6.14	0.9351
3	5.96	1.14	5.85	0.9517
4	7.04	0.05	7.04	0.9264
5	7.29	0.72	7.26	0.9015
6	6.77	−0.02	6.77	0.9134

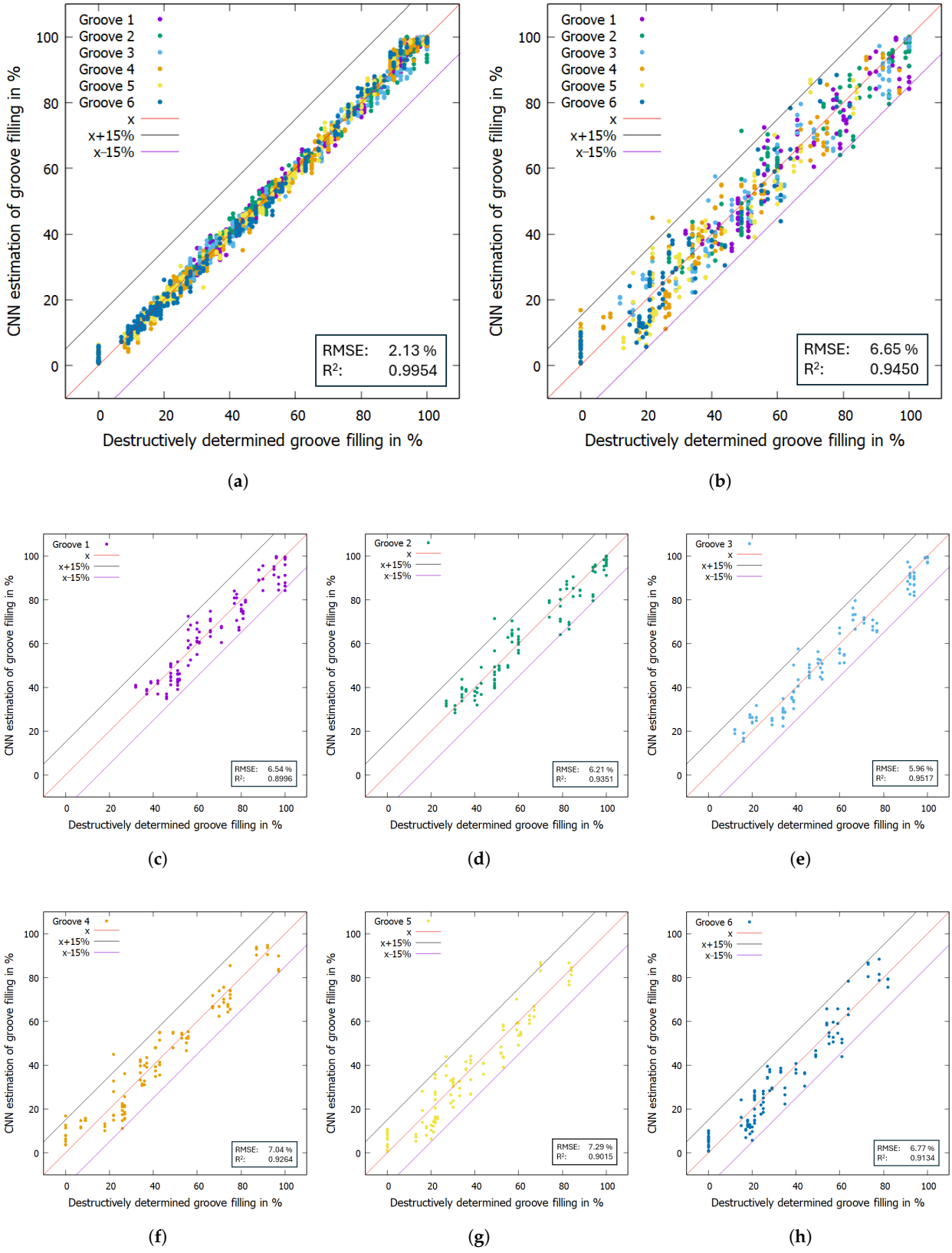


Figure 13. Comparison of the destructively determined groove fill levels and the output of the neural network: (**a**) training data; (**b**) test data; (**c**) test data—only first grooves; (**d**) test data—only second grooves; (**e**) test data—only third grooves; (**f**) test data—only fourth grooves; (**g**) test data—only fifth grooves; and (**h**) test data—only sixth grooves.

All B-scans used for training and testing contain realistic measurement disturbances such as speckle, small coupling variations, and minor geometric misalignments. No explicit denoising was applied, so the reported RMSE of approximately 7% already reflects the influence of these effects under typical immersion-tank conditions. In the available data, we did not observe systematic failure patterns that could be clearly attributed to noise spikes or isolated artifacts. However, extreme situations such as the complete loss of coupling or strong saturation effects are not represented in the present dataset and, therefore, remain outside the validated operating range of the method.

One of the reasons for using a neural network to determine the degree of filling based on ultrasound data was that the neural network would not only use the first echo to determine the degree of filling but would also take subsequent echoes into account, thus enabling a more precise determination of the degree of filling. To verify that the neural network actually "looks" also at later echoes, the explainable AI technique "guided Grad-CAM" was used [28] to visualize the regions in the input images (B-Scans) that are most decisive for the network's decision. Figure 14 shows an example of a B-Scan (a), together with the saliency map determined by Grad-CAM (b), as well as both overlaid (c). It is clear to see that CNN not only uses later echoes but also even focuses on them.

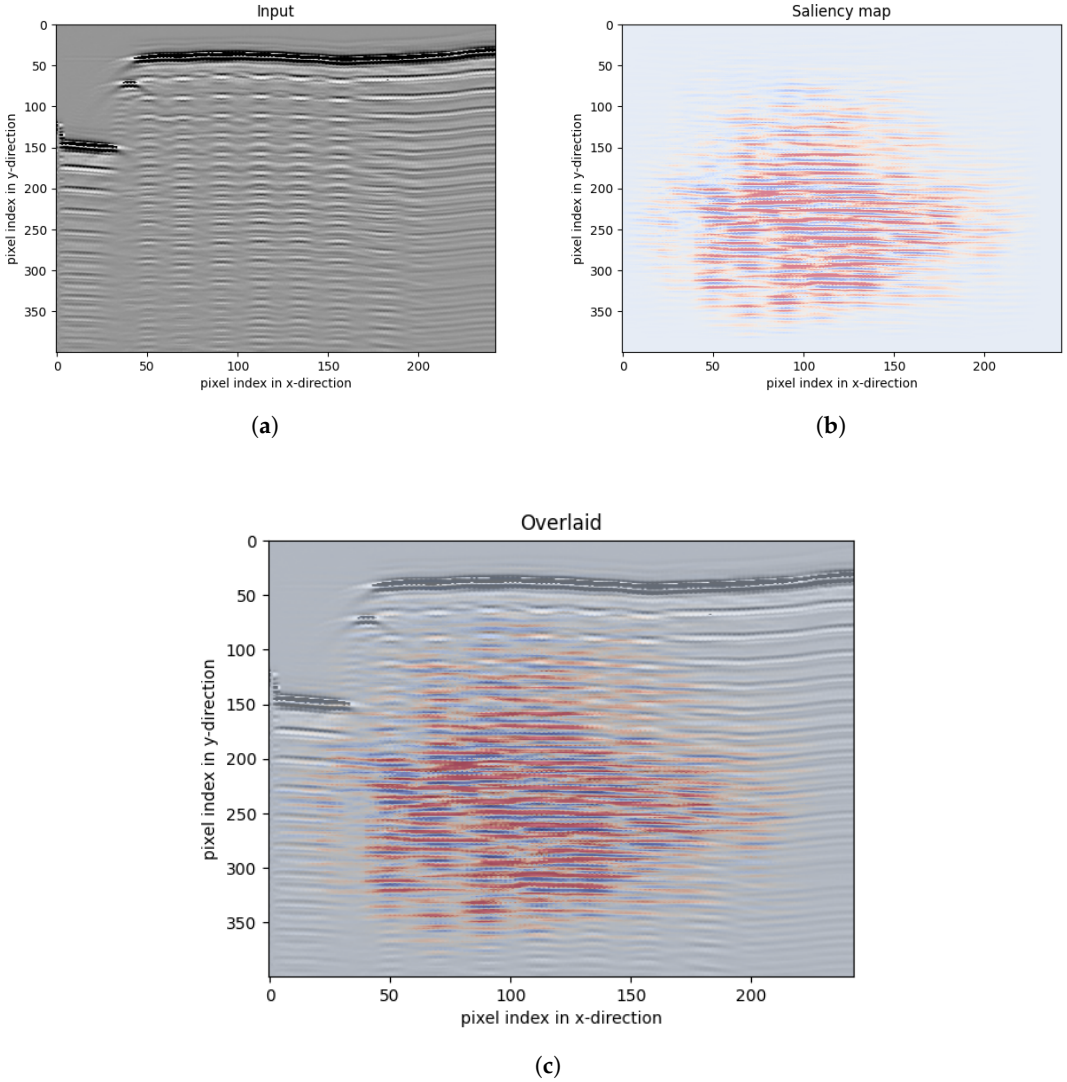


Figure 14. Visualization of Grad-CAM results for a BScan: (**a**) Input BScan. (**b**) Saliency map for this BScan. (**c**) BScan and saliency map overlaid.

In addition, the CNN-based evaluation compares favorably to the manual single-probe time-of-flight analysis discussed in Section 3. While the largest deviation in Table 1 lies in the order of magnitude of the groove heights themselves at certain probe positions, the CNN predictions remain within about 7% of the filling level across all grooves in the test set. Given that both manual and conceivable automated time-of-flight approaches would rely on a restricted number of hand-crafted timing features and explicit rules for echo selection, this underlines the benefit of exploiting the complete B-scan and multiple echo trains through data-driven learning instead of relying on hand-picked transit times and manually designed feature pipelines.

5. Conclusions

5.1. Summary

This work presented a nondestructive method to determine the filling level of six annular grooves with a cross section of 1 mm × 0.25 mm in titanium pipe-fitting connections. High-frequency phased array ultrasound at 20 MHz was used in an immersion setup to record raw B-scans from inside the pipe. Groove-wise filling levels between 0 percent and 100 percent were regressed by a convolutional neural network that was trained on B-scans and destructive reference measurements from micrographs. X-ray computed tomography could not provide sufficiently accurate groove filling levels due to a limited voxel resolution and the similar X-ray attenuation of pipe and fitting and was, therefore, used only as a reference for selected specimens. On a held-out test set of 36 B-scans the model achieved a root mean square error of about 7 percent of the filling level across all grooves. A guided Grad-CAM analysis showed that the network systematically exploits later echoes in the B-scan, which confirms that data-driven processing can use information that is difficult to capture with manual time of flight rules and with rule-based time of flight pipelines that rely on a small set of hand-crafted echo timing features.

5.2. Limitations and Applicability

The present study involved several limitations. The model was trained and evaluated on a single combination of pipe and fitting materials and on one specific groove geometry, so its applicability to other titanium alloys, wall thicknesses, or groove shapes has not yet been demonstrated. All measurements were carried out in a water bath with good coupling, while dry coupling, strong probe misalignment, and pronounced surface roughness were not investigated. The dataset comprises 25 joints and 150 B-scans and thus does not yet cover the full range of manufacturing variability that would be expected in production. In addition, training and test data were acquired on the same inspection system, so domain shifts such as different probes, electronics, or coupling media have not been assessed. Consequently, the current results should be regarded as a proof of concept for this specific configuration, rather than a fully validated industrial solution.

5.3. Outlook

Future work should extend the dataset to more joints, groove geometries, and material combinations and should incorporate explicit uncertainty quantification for each predicted filling level, for example, by Monte Carlo dropout or deep ensembles [23,29]. Physics-informed neural networks that embed basic ultrasonic propagation constraints into the loss function offer a promising way to improve generalization from limited data [30]. In parallel, the physics-based simulation of B-scans for different groove fillings could be used to generate synthetic training data and to study robustness with respect to noise, misalignment, and coupling variations. Finally, the integration of the proposed approach

into an inline capable inspection system with real-time inference is an important step toward industrial deployment in aerospace pipe production.

Author Contributions: Conceptualization, N.B. and J.P.-S.; methodology, K.J., N.B. and J.P.-S.; writing—original draft preparation, K.J. and B.S.; writing—review and editing, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: The funding of this investigation within the LuFo project "AutoFit" (FKZ: 20W1905D) as part of the 'National Civil Aviation Research Program LuFo VI-1, Program Line (C) Technology' by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) is gratefully acknowledged.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jacob, K.; Straß, B. *Hauptarbeitspaket 2–NDT-Verfahren (Einrollen)*; Technische Informationsbibliothek: Hannover, Germany, 2024. [CrossRef]
2. Verein Deutscher Ingenieure (VDI). *Computertomografie in der Dimensionellen Messtechnik–Grundlagen und Definitionen*; VDI/VDE 2630 Blatt 1.1; Verein Deutscher Ingenieure (VDI): Düsseldorf, Germany, 2009.
3. Beine, C.; Boller, C.; Netzelmann, U.; Porsch, F.; Venkat, R.; Schulze, M.; Bulavinov, A.; Heuer, H. NDT for CFRP Aeronautical Components a Comparative Study. In Proceedings of the 2nd International Symposium on NDT in Aerospace, Hamburg, Germany, 22–24 November 2010.
4. Maisl, M.; Schorr, C.; Porsch, F.; Haßler, U. Computerlaminographie, Grundlagen und technische Umsetzung. In Proceedings of the Industrielle Computertomografie: Zerstörungsfreie Bauteilprüfung, 3D-Materialcharakterisierung und Geometriebestimmung, Wels, Austria, 27–29 September 2010; pp. 261–266.
5. Straß, B.; Conrad, C.; Wolter, B. Production integrated nondestructive testing of composite materials and material compounds—An overview. *Iop Conf. Ser. Mater. Sci. Eng.* **2017**, *118*, 012017. [CrossRef]
6. Netzelmann, U. Nondestructive testing of ceramic-metal joints by high-frequency ultrasound techniques. In Proceedings of the Joining Ceramics, Glass and Metal, Königswinter, Germany, 17–19 May 1993; pp. 168–175.
7. Ramanan, S.V.; Bulavinov, A.; Pudovikov, S.; Boller, C.; Wenzel, T. Quantitative non-destructive evaluation of CFRP components by sampling phased array. In Proceedings of the 2nd International Symposium on NDT in Aerospace, Hamburg, Germany, 22–24 November 2010.
8. Siljama, O.; Koskinen, T.; Jessen-Juhler, O.; Virkkunen, I. Automated flaw detection in multi-channel phased array ultrasonic data using machine learning. *J. Nondestruct. Eval.* **2021**, *40*, 67. [CrossRef]
9. Virkkunen, I.; Koskinen, T.; Jessen-Juhler, O.; Rinta-Aho, J. Augmented ultrasonic data for machine learning. *J. Nondestruct. Eval.* **2021**, *40*, 4. [CrossRef]
10. Jia, Y.; Rakhmatov, D. Crack Defect Characterization Using Raw Channel Data and DNN-Based Classifier. In Proceedings of the 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS), Taipei, Taiwan, 22–26 September 2024; pp. 1–4.
11. Pyle, R.J.; Bevan, R.L.; Hughes, R.R.; Rachev, R.K.; Ali, A.A.S.; Wilcox, P.D. Deep learning for ultrasonic crack characterization in NDE. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2020**, *68*, 1854–1865. [CrossRef] [PubMed]
12. Bai, L.; Le Bourdais, F.; Miorelli, R.; Calmon, P.; Velichko, A.; Drinkwater, B.W. Ultrasonic defect characterization using the scattering matrix: A performance comparison study of Bayesian inversion and machine learning schemas. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2021**, *68*, 3143–3155. [CrossRef] [PubMed]
13. Shi, Y.; Xu, W.; Zhang, J.; Li, X. Automated classification of ultrasonic signal via a convolutional neural network. *Appl. Sci.* **2022**, *12*, 4179. [CrossRef]
14. Latête, T.; Gauthier, B.; Belanger, P. Towards using convolutional neural network to locate, identify and size defects in phased array ultrasonic testing. *Ultrasonics* **2021**, *115*, 106436. [CrossRef] [PubMed]
15. Naddaf-Sh, A.M.; Baburao, V.S.; Zargarzadeh, H. Automated Weld defect detection in industrial ultrasonic B-scan images using deep learning. *NDT* **2024**, *2*, 108–127. [CrossRef]

16. Krautkrämer, J.; Krautkrämer, H. *Ultrasonic Testing of Materials*; Springer: Berlin/Heidelberg, Germany, 1990.
17. Drinkwater, B.W.; Wilcox, P.D. Ultrasonic arrays for non destructive evaluation—A review. *Ndt E Int.* **2006**, *39*, 525–541. [CrossRef]
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [CrossRef]
21. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* **2017**, arXiv:1706.05098. [CrossRef]
22. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017. [CrossRef]
23. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning, PMLR 48, New York, NY, USA, 19–24 June 2016; pp. 1050–1059. [CrossRef]
24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015. [CrossRef]
25. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
26. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), PMLR 37, Lille, France, 7–9 July 2015; pp. 448–456. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
28. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
29. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6402–6413. [CrossRef]
30. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]