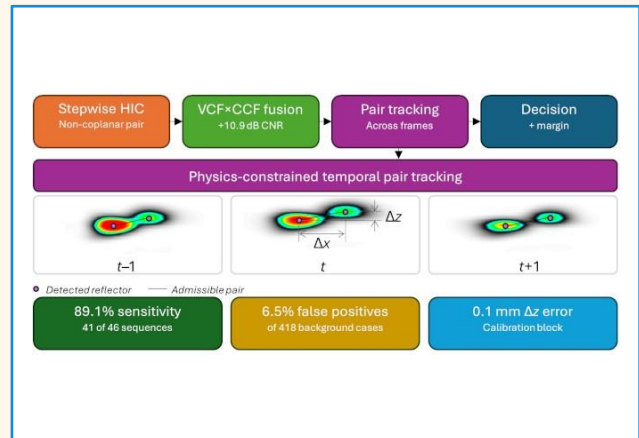


Detection and Characterization of Hydrogen-Induced Stepwise Cracks Using Multi-Plane Wave Imaging, Coherence Factor Fusion and Physics-Constrained Temporal Pair Tracking

Vykintas Samaitis, *Member, IEEE*, Renaldas Raišutis, *Member, IEEE*, Audrius Stravinskas, Aadhik Asokkumar

Abstract — Stepwise hydrogen-induced cracking (HIC) in carbon-steel components rarely appears as a single resolvable crack, instead it produces compact, non-coplanar reflectors distributed through the wall thickness. This study presents an automated, physically interpretable ultrasonic framework that detects this indirect signature directly. Its core is a deterministic, physics-constrained temporal pair-tracking detector that forms candidate non-coplanar reflector pairs within each scan frame, links physically admissible pairs across neighbouring frames and accepts only tracks satisfying predefined score, persistence and geometric criteria. Unlike data-driven classifiers, every decision is traceable to explicit geometric constraints and is accompanied by an interpretable decision margin. As a front end, plane wave imaging data were reconstructed with the total focusing method and several coherence-factor combinations were evaluated. Although SCF×CCF gave the highest CNR and lowest API, VCF×CCF was selected for its compact reflector-pair morphology and best detector compatibility, while still increasing CNR by 10.9 dB and reducing API from 7.77 to 0.47 relative to conventional PWI/TFM. The targeted signature was grounded in metallographic observation of stepwise cracking and in controlled reflector geometries in a calibration block, where calibration-compatible indications were detected with an absolute vertical-separation error of 0.1 mm. Relative to the expert NDT reference standard, the detector identified 41 of 46 positive sequences (89.1% sensitivity) and produced a 6.5% false-positive rate across 418 strict-background sequences. Ablation and robustness analyses confirmed that performance arose from the combined score, persistence and geometry criteria rather than any single threshold. The proposed framework supports identification and prioritization of rare stepwise-cracking indications for further operator review.



Index Terms — coherence factor fusion; hydrogen-induced cracking; plane wave imaging; stepwise cracking; temporal pair tracking; total focusing method; ultrasonic non-destructive testing.

I. INTRODUCTION

The pursuit of net-zero and carbon neutrality targets has intensified efforts to identify viable alternatives to conventional natural gas. Among these, hydrogen (H_2) has emerged as a promising energy carrier due to its potential to deliver energy without associated carbon dioxide (CO_2) emissions. To address current challenges related to large-scale hydrogen transportation over long distances, there is growing

interest in repurposing existing natural gas pipeline infrastructure for hydrogen blending and distribution [1]. However, most existing gas pipeline networks were originally designed for fossil gas mixtures dominated by methane (CH_4), with minimal consideration for hydrogen (H_2) presence [2]. Consequently, the materials, components and operational parameters of these systems were optimized and validated for environments containing very low concentrations of H_2 . Numerous studies have identified key challenges in adapting

This work was supported by the Research Foundation of the Research Council of Lithuania under the project CURIOUS. Grant no. S-MIP-24-45.

V. S. is with the Department of Ultrasound Research Institute, Kaunas University of Technology, K.Barsausko st. 59, LT-51423, Kaunas, Lithuania (e-mail: vykintas.samaitis@ktu.lt).

R. R. is with the Department of Ultrasound Research Institute and department of Electrical Power systems, Faculty of Electrical and Electronics engineering, Kaunas University of Technology,

K.Barsausko st. 59, LT-51423, Kaunas, Lithuania (e-mail: renaldas.raisutis@ktu.lt).

A. S. is with the Department of Ultrasound Research Institute, Kaunas University of Technology, K.Barsausko st. 59, LT-51423, Kaunas, Lithuania (e-mail: audrius.stravinskas@ktu.lt).

A. A. is with the Department of Ultrasound Research Institute, Kaunas University of Technology, K.Barsausko st. 59, LT-51423, Kaunas, Lithuania (e-mail: aadhik.asokkumar@ktu.lt).

Highlights

- A physics-constrained temporal pair-tracking detector was developed for stepwise-HIC-like signatures in coherence-factor-fused PWI/TFM images.
- The detector achieved 89.1% sensitivity and 6.5% false-positive rate on 46 HIC-positive and 418 strict-background sequences.
- The approach provides an interpretable NDT framework for flagging rare, high-importance stepwise HIC indications for targeted operator review.

current infrastructure, emphasizing the need to strengthen the hydrogen compatibility of pipeline materials, ensure hydrogen suitability and maintain operational safety [3-6]. Hydrogen-induced cracking (HIC) is a common degradation mechanism in low-to medium-strength carbon steel pipelines exposed to high-pressure gaseous hydrogen environments, where the atomic hydrogen permeates into the steel matrix and accumulates at microstructural trap sites such as non-metallic inclusions, voids, grain boundaries and laminations [6,7]. At these trap sites, hydrogen atoms can recombine to form molecular hydrogen, generating localized high internal pressures. According to the internal pressure theory, when the pressure of molecular hydrogen between an inclusion and the surrounding metal matrix exceeds a critical threshold, it can initiate microcracks, ultimately leading to HIC propagation, especially at the girth weld zone [8-10]. Once a crack initiates, continued ingress of atomic hydrogen leads to the accumulation of molecular hydrogen at the crack tip, promoting self-propagation of the crack into the material [11]. HIC typically propagate parallel to the steel surface, or in the plane of the material, a behaviour often associated with the rolling process during steel manufacturing. These parallel, non-coplanar cracks can connect through their crack tips, forming what is known as stepwise cracking (SWC). When stepwise cracking causes a significant reduction in the effective wall thickness of the pipe, it can ultimately lead to pipeline failure [12].

Several ultrasonic non-destructive testing (UT) techniques, including conventional UT, time-of-flight diffraction (TOFD), phased array ultrasonic testing (PAUT), have been employed for the detection of hydrogen-induced damage [13-15]. Conventional UT approaches often utilize methods such as the Advanced Velocity Ratio (AVR) and Advanced Ultrasonic Backscatter Technique (AUBT) [16,17]. The AVR technique relies on the measurement of longitudinal-to-shear wave velocity ratios, while AUBT evaluates the amplitude of backscattered signals caused by HIC-related microstructural changes, or analyses signal attenuation in the frequency domain. TOFD and its variant, the Ultra Low Angle TOFD (TULA), operate based on the detection of diffracted waves originating from crack tips, using a pair of acoustically isolated angle-beam probes positioned opposite each other. Despite their demonstrated capabilities, these techniques face limitations, including reduced sensitivity to damage near sample boundaries, the need for significant crack propagation prior to detection, pointwise inspection and robustness challenges in field applications [18-20]. One of the primary limitations of conventional NDT techniques is their insufficient resolution, particularly in detecting early-stage hydrogen-induced cracking (HIC) and stepwise cracking. The emergence of ultrasonic phased array technology, combined with

advanced data acquisition methods such as Full Matrix Capture (FMC) and Multi-Plane Wave Imaging (MPWI), along with post-processing techniques like Delay-and-Sum (DAS), has enabled the generation of fully focused volumetric ultrasonic images in the near field by computing the pixel intensity within a region of interest (ROI) using appropriately delayed signals [21-24]. Recent enhancements to TFM post-processing involve image weighting using coherence factors, which leverage the instantaneous phase information of the received signals which allows for amplitude-independent volumetric reconstructions where pixel intensity reflects the degree of phase coherence across the aperture [25,26]. Various coherence-based weighting techniques have been proposed in the literature, including Phase Coherence Factor (PCF), Sign Coherence Factor (SCF), Circular Coherence Factor (CCF) and Vector Coherence Factor (VCF) [27-32]. These methods have demonstrated significant improvements in lateral resolution, side-lobe suppression and enhancement of diffracted signals from crack tips. They have been successfully applied to the detection of high-temperature hydrogen attack (HTHA), fatigue cracks and for the estimation of crack orientation [21,30,31,33]. While coherence-based weighting factors and their variants have shown strong potential for distinguishing voids and fissures from structural noise, the detection and characterization of stepwise cracking arising from the linkage of non-coplanar defects remains a major challenge due to limited contrast and resolution. In conventional TFM imaging, the inherent resolution limitations often result in closely spaced, non-coplanar reflectors being interpreted as a single defect. However, from a structural integrity standpoint, stepwise crack formations characterized by linked micro-defects distributed through the thickness pose a substantially higher safety risk than isolated scatterers. This is because linked non-coplanar HIC blisters can reduce the effective load-bearing wall thickness of the component, even when the individual crack segments are not fully resolved as a continuous crack in the ultrasonic image. In addition, the short vertical or inclined ligaments that connect adjacent HIC blisters may produce weak or unfavourably oriented ultrasonic responses, making direct detection of the interlinking path difficult. Therefore, recognizing the indirect ultrasonic signature of stepwise cracking, namely stable pairs of non-coplanar reflectors separated in depth and lateral position, is important for reliable NDT assessment of HIC-damaged components. Detecting such defects therefore requires further improvements in imaging resolution and contrast. In parallel, machine-learning and deep-learning approaches have recently been applied to ultrasonic crack detection, sizing and orientation estimation, achieving high accuracy on diffraction and image data [34-36]. However, these data-driven methods typically

require large annotated training sets and offer limited transparency into why a given indication is flagged, which is problematic for rare, safety-critical stepwise-HIC indications that demand targeted operator review. For this problem, automated analysis should therefore remain physically interpretable, tracing each decision to explicit, geometry-based criteria rather than to a black-box classifier.

In this study, we propose a physics-constrained temporal pair-tracking approach for detecting stepwise hydrogen-induced cracking in metallic heat-exchanger shell structures exposed to hydrogen-rich environments. The central idea is to target the indirect ultrasonic signature of stepwise cracking, two compact, non-coplanar reflectors separated in depth and lateral position that persist over neighbouring axial scan frames, rather than attempting to image a single continuous crack, which is rarely resolvable for through-thickness linked HIC blisters. A deterministic detector first extracts compact reflector candidates from each frame, forms physically admissible non-coplanar reflector pairs, links these pairs across neighbouring scan frames and accepts a sequence only when the resulting temporal track satisfies predefined score, persistence, amplitude-balance and geometric criteria. In contrast to data-driven classifiers [34-36], every acceptance decision is traceable to explicit physical constraints and is accompanied by an interpretable decision margin. As a front end to this detector, coherence-factor fusion is used to suppress structural noise and sharpen closely spaced, non-coplanar reflectors prior to detection. Multiplicative combination of coherence information has been reported previously [32,33]; the contribution here is not the multiplication operation itself, but the systematic evaluation of all pairwise fusions of VCF, CCF, SCF and PCF and the identification of $VCF \times CCF$ as the best-balanced fusion for compact non-coplanar HIC reflectors. The physical basis of the targeted signature is established from metallographic observation of stepwise HIC in one sample and from controlled non-coplanar reflector geometries in an artificial calibration block. The proposed approach is then evaluated on real HIC-positive and strict-background image sequences extracted from naturally damaged heat-exchanger shell samples. The framework is intended to support expert NDT interpretation by flagging rare but structurally important stepwise-HIC-like indications for further operator review.

II. MATERIALS AND METHODS

A. Object under inspection

In this study, two segments of in-service carbon-steel heat-exchanger shells containing naturally developed hydrogen-induced cracking (HIC) damage were investigated. Hereafter, the two shell sections are referred to as S4 and S5 according to their part numbers. Both shell sections had an outer diameter of 650 mm and a wall thickness of 46 mm. The dimensions of S4 were 540 mm $\times$ 556 mm, while S5 measured 425 mm $\times$ 535 mm. The S4 section contained both axial and circumferential V-groove welds located along the edges of the sample, as shown in Figure 1(a). The S5 section contained a circumferential V-groove weld located approximately at the centre of the sample in the axial direction, as shown in Figure 1(b). The detailed spatial distributions of HIC damage in these samples were previously reported in [20]. In S4, the HIC

distribution was non-uniform, with a higher concentration of indications near the intersection of the axial and circumferential weld regions. In S5, the HIC distribution was more uniform on the near side of the weld and more non-uniform on the far side. In both samples, the majority of HIC indications were located within the depth range of 41–46 mm from the outer surface.

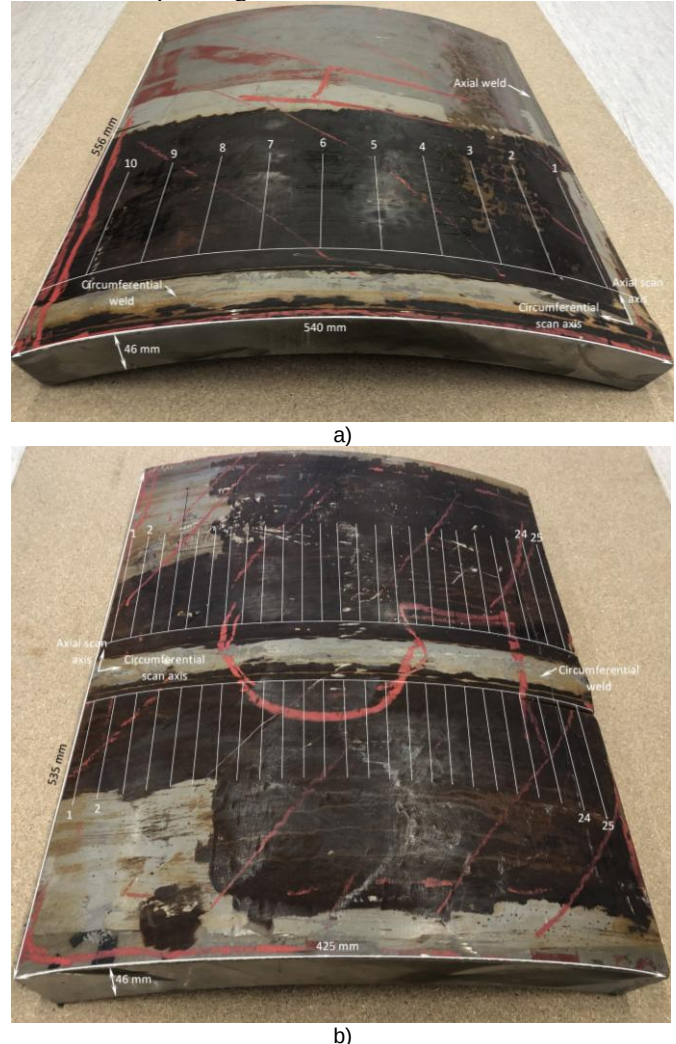


Fig. 1. The segments of the carbon steel heat exchanger shells used for inspection of HIC damage: a) sample S4 and b) sample S5. The white lines indicate circumferential scan positions.

Metallographic analysis was performed on sample S5 and revealed the presence of different HIC morphologies. In regions close to the weld, isolated HIC blisters were observed. In regions located further from the weld, a stepwise cracking case associated with blister interlinking was identified. The maximum measured HIC penetration depth was 8.8 mm from the inner surface of the shell wall. The metallographically confirmed stepwise morphology in S5 was used as the physical reference for defining the characteristic ultrasonic signature targeted by the proposed detection approach. The metallographic image of the stepwise defect and its geometrical interpretation is presented in Section II.G.

B. Description of coherence factors

The classical Total Focusing Method (TFM) is based on the Full Matrix Capture (FMC) acquisition technique and performs

discrete image reconstruction within a defined ROI by summing time-delayed signals from all transmitter-receiver element pairs. This summation aims to estimate the scattering amplitude at each image point (x, z) , effectively synthesizing a fully focused image across the entire field:

$$I(x, z) = \frac{1}{N \cdot L} \sum_{k=1}^L \sum_{i=1}^N r_{k,i} \left(\tau_{k,i}(x, z) \right) \quad (1)$$

where $\tau_{k,i}(x, z)$ is the time of flight from the k^{th} transmitter, to the pixel located at coordinates (x, z) and back to the i^{th} receiver. A key limitation of conventional TFM, however lies in its inconsistent image intensity response across different types of reflectors. Specifically, omnidirectional scatterers often produce scattered signals with relatively low amplitude leading to weaker coherent summation. In contrast, the specular reflections from smooth surfaces, such as backwalls, sum to large values not necessarily due to improved coherence, but rather because their reflection amplitudes are orders of magnitude greater than those from small scatterers. This discrepancy results in image imbalance, wherein weak yet critical scattering features are obscured or underrepresented in the final image due to the overwhelming influence of high-amplitude specular reflections [31]. One straightforward strategy is to scale the TFM image, thereby suppressing amplitude-driven dominance and enhancing the relative visibility of phase-coherent scatterers. This can be achieved by measuring the phase dispersion of the delayed signals by using various coherence factors such as PCF, CCF, SCF and VCF [28,37].

The phase coherence factor is calculated from the instantaneous phase considering that phase is linear within the $(-\pi, \pi]$ [28,38]:

$$PCF(x, z) = \max \left(0, 1 - \frac{\text{std}(\varphi_{k,i}(x, z))}{\frac{\pi}{\sqrt{3}}} \right), \quad (2)$$

where $\text{std}(\cdot)$ is the standard deviation operator, φ is the instantaneous phase obtained using Hilbert transform operator, $\frac{\pi}{\sqrt{3}}$ is the normalization coefficient defining standard deviation of the uniform distribution. The sign coherence factor uses the sign of the signal for phase discretization with single bit [28,29]:

$$SCF(x, z) = 1 - \sqrt{1 - \left[\frac{1}{N \cdot L} \sum_{k=1}^L \sum_{i=1}^N \text{sign} \left(r_{k,i} \left(\tau_{k,i}(x, z) \right) \right) \right]^2}, \quad (3)$$

where $\text{sign}(\cdot)$ is the sign operator. The circular coherence factor considers that phase has a continuous circular distribution in $(0, 2\pi]$ [28,29]:

$$CCF(x, z) = 1 - \sqrt{\text{var} \left(\cos(\varphi_{k,i}(x, z)) \right) + \text{var} \left(\sin(\varphi_{k,i}(x, z)) \right)}, \quad (4)$$

where $\text{var}(\cdot)$ is the variance operator. Finally the vector coherence factor is defined as set of unit vectors in the complex plane where the modulus of the vector sum measures the phase concentration [28]:

$$VCF(x, z) = \frac{1}{N \cdot L} \sqrt{\left(\sum_{k=1}^L \sum_{i=1}^N \frac{\text{Re}(s_{k,i}(x, z))}{|s_{k,i}(x, z)|} \right)^2 + \left(\sum_{k=1}^L \sum_{i=1}^N \frac{\text{Im}(s_{k,i}(x, z))}{|s_{k,i}(x, z)|} \right)^2}, \quad (5)$$

where $s_{k,i}(x, z)$ is the complex analytic representation of delayed signals, $|\cdot|$ denotes the modulus operator, Re and Im denote the real and imaginary parts of the complex analytic signal.

Each coherence factor offers distinct advantages and limitations in the context of ultrasonic imaging. The SCF is computationally efficient, making it well-suited for real-time or

resource-constrained applications. However, due to the binarization of phase information, SCF is susceptible to high-frequency artifacts and often necessitates the application of smoothing filters to suppress spurious responses caused by abrupt phase transitions. In contrast, the CCF inherently yields a smoother coherence map, owing to its statistical treatment of phase variance. Despite this, CCF is computationally more demanding due to the required variance calculation and is sensitive to small phase misalignments, which can lead to a distorted representation of discontinuities or cracks in certain scenarios [27,29]. As a compromise between computational cost and robustness, the VCF has been proposed. It demonstrates greater tolerance to minor phase errors and offers improved computational efficiency compared to CCF, while still preserving much of the spatial coherence information necessary for accurate defect visualization [29]. In order to obtain the metric that encompasses the benefits of several weighting factors, such as smoother weighting function, increased tolerance to phase errors and better reduction of the noise floor we explored fusion of several coherence factors. This fusion approach aims to enhance imaging contrast while preserving high spatial resolution of defect features and is described in the following section.

C. Coherence factor fusion

In this study, we explore fusion of the VCF, CCF, SCF and PCF to derive a composite beamformer-output weighting function that simultaneously enhances image resolution and suppresses background noise. Multiplicative combination of coherence information has previously been used to improve crack visibility, for example through sign-multiply and angle-weighted sign coherence factors [32-33]. Here, the fusion is performed between all possible pairwise variants using a straightforward pixel-wise multiplication of the individual coherence maps and the resulting combinations are then compared systematically to identify the most suitable fusion for compact non-coplanar reflectors (section III.A). These fused maps aim to emphasize reflectors that are consistently assigned high coherence by both constituent factors and to suppress noisy features that are inconsistent between them. The pixel-wise multiplication fusion can be expressed as:

$$F_{i,j}(x, z) = C_i(x, z) \cdot C_j(x, z), \quad (6)$$

where $C_i(x, z)$ and $C_j(x, z)$ represent the coherence weighting factors being fused, with $i < j$ and $i, j \in \{1 \dots N\}$, where N is the total number of considered coherence factors. While alternative fusion formulations could be implemented to compensate for individual factor distributions, direct multiplicative fusion is chosen here for its mathematical properties as a non-linear spatial AND gate. Because each individual coherence factor map is strictly bounded within the interval $[0, 1]$, direct multiplication acts as a severe penalty function that exponentially suppresses background noise regions where either constituent factor exhibits poor phase concentration. Furthermore, because the proposed framework utilizes an adaptive, variance-maximizing Otsu threshold immediately following the fusion stage, any global variations in the composite dynamic range are automatically compensated for on a frame-by-frame basis. In this study, four coherence factors VCF, CCF, SCF and PCF are considered, resulting in six

pairwise fused images: VCF×SCF, VCF×CCF, VCF×PCF, SCF×CCF, SCF×PCF and CCF×PCF. Based on the performance evaluation of different combinations of the coherence factors, only one of them will later be selected for further analysis which is demonstrated in section III.A.

D. Image threshold definition

To further suppress the noise floor and enhance image contrast, a thresholding step was applied following the fusion of coherence factors and weighting of the PWI image, as proposed in [26]. In this study, we used a moderate-level variance-based threshold (Otsu's method), which finds the threshold that best separates foreground and background by maximizing the difference between class means, weighted by their probabilities [39]:

$$t^* = \arg \max_t \sigma_B^2(t), \quad (7)$$

$$\sigma_B^2(t) = \omega_0(t) \cdot \omega_1(t) \cdot (\mu_0(t) - \mu_1(t))^2, \quad (8)$$

where $\omega_0(t)$ and $\omega_1(t)$ are the class probabilities for background and foreground, $\mu_0(t)$ and $\mu_1(t)$ are the respective class means. The thresholded image is then provided as:

$$I_T(x, z) = \begin{cases} I(x, z), & \text{if } I(x, z) \geq t^* \\ 0, & \text{if } I(x, z) < t^* \end{cases} \quad (9)$$

E. Quantitative imaging performance evaluation metrics

To quantitatively assess the contrast between the HIC-induced reflections and the surrounding structural noise and to evaluate the improvement proposed by coherence factor fusion, the contrast-to-noise ratio (CNR) was employed as the evaluation metric. CNR was computed as the ratio of the absolute difference between the mean signal amplitude within the damage region and the mean noise level, normalized by the standard deviation of the noise and expressed in decibels (dB):

$$\text{CNR} = 20 \log_{10} \left(\frac{|\mu_{\text{flaw}} - \mu_{\text{background}}|}{\sigma_{\text{background}}} \right), \quad (10)$$

where μ_{flaw} and $\mu_{\text{background}}$ – are the average intensity of the signal in defect area and background area respectively, $\sigma_{\text{background}}$ – is the standard deviation of the noise. This metric captures the average distinguishability between signal and noise by incorporating the statistical spread of the noise and mitigating the influence of isolated high-intensity peaks through the use of mean signal amplitude. This specific metric is referenced in ASTM standards and has been demonstrated to offer increased robustness compared to other commonly used definitions of CNR [40,41]. While CNR does not account for spatial resolution and remains sensitive to ROI selection, the Array Performance Index (API) was employed as a complementary metric to evaluate the imaging system's ability to resolve small scatterers. The API is defined as the area of the ultrasonic image exceeding a –6 dB threshold, normalized by the square of the wavelength as [24,42]:

$$\text{API} = \left(\frac{S_{-6\text{dB}}}{\lambda^2} \right), \quad (11)$$

where $S_{-6\text{dB}}$ is the area of image in which amplitude is higher than highest amplitude at –6 dB, λ – is the wavelength in steel. API is a dimensionless measure that decreases with the increase of image resolution.

Because CNR and API do not directly quantify the morphology of the detected reflector response, a connected-component morphology analysis was also used to support the

fusion-method comparison. The analysis was performed inside the same signal region used for the CNR and API estimation. To provide a standardized measure of the high-intensity reflector core, the signal region was thresholded at –6 dB relative to its local maximum, consistent with the API definition. This threshold was used only for morphology measurement and did not alter the detector input images. Connected components were identified using 8-connectivity. For each component m , the physical area A_m and perimeter P_m were estimated, and a compactness index $C_m = 4\pi A_m / P_m^2$ was calculated. A value approaching one indicates a compact, approximately circular component, whereas lower values indicate a more elongated, irregular or fragmented component boundary. Because the targeted stepwise signature consists of two reflector components, the main reflector pair was selected from the thresholded connected components using the same frame-level geometric admissibility limits as the detector. Among component pairs satisfying $\Delta x_{\min} \leq \Delta x \leq \Delta x_{\max}$ and $\Delta z_{\min} \leq \Delta z \leq \Delta z_{\max}$, the pair with the largest combined thresholded area was treated as the main reflector pair. If no component pair satisfied these limits, no detector-compatible main pair was assigned for that fusion image. The area-weighted compactness of this selected main pair, the total thresholded area, the number of connected components and the satellite area fraction outside the selected main pair were then used as morphology descriptors.

F. Proposed stepwise cracking detection approach

The proposed stepwise cracking detection procedure was formulated as a physics-constrained temporal pair-tracking approach. The method was designed to identify the characteristic ultrasonic response of stepwise hydrogen-induced cracking, where two compact, non-coplanar reflectors are expected to appear within a local region and to remain detectable over neighbouring axial scan positions. Therefore, the detection procedure did not rely on image amplitude alone, but combined local reflector extraction, geometrically admissible reflector-pair formation, temporal linking and final track-level acceptance criteria. Let $I_t(x, z)$ denote the reconstructed and coherence-factor-fused image at axial frame t , where x is the lateral coordinate and z is the depth coordinate. A local sequence of T neighbouring frames is denoted as $\mathbf{I} = \{I_t(x, z)\}$ for $t=1, 2, \dots, T$. The proposed detector operates on this image sequence and returns a binary sequence-level decision indicating whether a stepwise crack-like reflector track is present.

For each frame, local reflector candidates were first extracted from the fused image. A mild spatial smoothing operation was applied to suppress pixel-scale fluctuations while preserving compact reflector responses. Local maxima were then searched inside the region of interest. To reduce the dependence on absolute amplitude and to account for frame-to-frame variations in the background level, candidate strength was evaluated using a robust local contrast measure. For frame t , the robust background level was estimated as:

$$\mu_t = \text{median}(I_t), \quad (12)$$

and the robust scale was estimated as:

$$\sigma_t = 1.4826 \cdot \text{median}(|I_t - \mu_t|). \quad (13)$$

The factor 1.4826 is the Gaussian consistency correction for the median absolute deviation, converting median absolute

deviation into an equivalent standard-deviation estimate under normally distributed background noise. For a candidate reflector located at (x_i, z_i) , the robust contrast was then defined as:

$$R_i^{(t)} = \frac{(|I_t(x_i, z_i) - \mu_t|)}{(\sigma_t + \varepsilon)}, \quad (14)$$

where ε is a small positive constant used to avoid division by zero. A candidate was retained only if it satisfied minimum absolute amplitude, minimum amplitude relative to the maximum value in the corresponding frame and minimum robust-contrast requirements. Small connected components were rejected and only a limited number of dominant maxima were retained from each component. This allowed secondary reflector responses to be preserved when they were spatially separated, while suppressing isolated noise peaks and low-amplitude satellite responses. Each retained reflector candidate in frame t was represented as:

$$c_i^{(t)} = (x_i^{(t)}, z_i^{(t)}, A_i^{(t)}), \quad (15)$$

where $A_i^{(t)} \in (0, 1]$ represents the candidate peak intensity normalized by the maximum pixel intensity in frame t . Candidate reflectors were then combined into possible stepwise-crack pairs. For a pair of candidates i and j in frame t , the lateral and vertical separations were calculated as:

$$\Delta x_{i,j}^{(t)} = |x_i^{(t)} - x_j^{(t)}|, \quad (16)$$

$$\Delta z_{i,j}^{(t)} = |z_i^{(t)} - z_j^{(t)}|, \quad (17)$$

A candidate pair was considered geometrically admissible only if $\Delta x_{\min} \leq \Delta x_{i,j}^{(t)} \leq \Delta x_{\max}$ and $\Delta z_{\min} \leq \Delta z_{i,j}^{(t)} \leq \Delta z_{\max}$. The lower bound on Δx rejected nearly vertically stacked responses, while the upper bound rejected pairs that were too laterally separated to represent a local stepwise morphology. Similarly, lower bound on Δz was used to prevent a single broadened reflector from being interpreted as a stepwise pair, whereas the upper bound restricted the detection to the expected vertical separation range of local stepwise cracking. An amplitude-balance term was introduced to suppress pairs in which one reflector dominated the response. For each admissible pair, this term was defined as:

$$B_{i,j}^{(t)} = \frac{\min(A_i^{(t)}, A_j^{(t)})}{\max(A_i^{(t)}, A_j^{(t)}) + \varepsilon}. \quad (18)$$

Only pairs satisfying $B_{i,j}^{(t)} \geq B_{\min}$ were retained. A frame-level pair score was then calculated as:

$$S_{i,j}^{(t)} = \sqrt{A_i^{(t)} A_j^{(t)}} \cdot B_{i,j}^{(t)} \cdot \exp\left(\frac{-\Delta x_{i,j}^{(t)}}{\lambda_x}\right), \quad (19)$$

where λ_x is a lateral decay constant. The first term favours pairs formed by two strong reflectors, the second term favours amplitude-balanced pairs and the exponential term penalizes large lateral separations within the accepted geometric range.

In the next stage, admissible pairs were linked across neighbouring axial frames to form temporal tracks. Each pair was described by its centre coordinate

$$x_{c_{i,j}}^{(t)} = \frac{x_i^{(t)} + x_j^{(t)}}{2}, \quad (20)$$

and by the depths of the upper and lower reflectors, denoted $z_u^{(t)}$ and $z_l^{(t)}$, respectively. A pair in frame t was allowed to connect to a pair in frame t' only if the frame gap and spatial changes were within prescribed tolerances:

$$|t' - t| \leq g_{\max}, \quad (21)$$

$$|x_c^{(t')} - x_c^{(t)}| \leq \delta x_c, \quad (22)$$

$$|z_u^{(t')} - z_u^{(t)}| \leq \delta z_u, \quad (23)$$

$$|z_l^{(t')} - z_l^{(t)}| \leq \delta z_l, \quad (24)$$

$$|\Delta z^{(t')} - \Delta z^{(t)}| \leq \delta \Delta z, \quad (25)$$

here, $g_{\max}$ is the maximum allowed frame gap, δx_c is the maximum allowed shift in pair-centre location, δz_u and δz_l are the maximum allowed shifts of the upper and lower reflector depths and $\delta \Delta z$ is the maximum allowed change in the vertical reflector separation. These constraints enforce the assumption that a real stepwise-crack response should persist and evolve smoothly over adjacent axial positions, rather than appearing as unrelated isolated reflector pairs. For each temporal track $\mathcal{T}$, the longest consecutive run length was denoted as $L(\mathcal{T})$. The maximum and median pair scores within the track for all pairs (i, j, t) belonging to $\mathcal{T}$ were defined as:

$$S_{\max}(\mathcal{T}) = \max[S_{i,j}^{(t)}], \quad (26)$$

$$S_{\text{med}}(\mathcal{T}) = \text{median}[S_{i,j}^{(t)}], \quad (27)$$

The median vertical and lateral separations of the track for all pairs (i, j, t) belonging to $\mathcal{T}$ were calculated as:

$$\Delta \tilde{x}(\mathcal{T}) = \text{median}[\Delta x_{i,j}^{(t)}], \quad (28)$$

$$\Delta \tilde{z}(\mathcal{T}) = \text{median}[\Delta z_{i,j}^{(t)}]. \quad (29)$$

A track was accepted as a stepwise-crack indication only if all final track-level criteria were satisfied: $L(\mathcal{T}) \geq L_{\min}$, $S_{\max}(\mathcal{T}) \geq S_{\max, \min}$, $S_{\text{med}}(\mathcal{T}) \geq S_{\text{med, min}}$, $\Delta \tilde{z}(\mathcal{T}) \leq \Delta \tilde{z}_{\max}$, $\Delta \tilde{x}_{\min} \leq \Delta \tilde{x}(\mathcal{T}) \leq \Delta \tilde{x}_{\max}$. Thus, a reflector pair had to be sufficiently persistent, sufficiently strong, sufficiently stable and geometrically consistent with local non-coplanar stepwise cracking. The sequence-level detector output was then defined as $D(\mathcal{J}) = 1$, if at least one track $\mathcal{T}$ satisfies all acceptance criteria and $D(\mathcal{J}) = 0$ otherwise. In addition to the binary decision, a normalized decision margin was calculated for each track. This margin provides an interpretable measure of the distance from the final multi-criterion decision boundary. Unlike deep-learning crack-detection and sizing methods [34-36], which can achieve high accuracy but provide limited insight into individual decisions and require substantial annotated training data, the proposed detector links every accepted track to explicit, physically meaningful criteria. For a track $\mathcal{T}$, the margin was defined as:

$$M(\mathcal{T}) = \min\left[\frac{(L(\mathcal{T}) - L_{\min})}{L_{\min}}, \frac{(S_{\max}(\mathcal{T}) - S_{\max, \min})}{S_{\max, \min}}, \frac{(S_{\text{med}}(\mathcal{T}) - S_{\text{med, min}})}{S_{\text{med, min}}}, \frac{(\Delta \tilde{z}_{\max} - \Delta \tilde{z}(\mathcal{T}))}{\Delta \tilde{z}_{\max}}, \frac{(\Delta \tilde{x}(\mathcal{T}) - \Delta \tilde{x}_{\min})}{\Delta \tilde{x}_{\min}}, \frac{(\Delta \tilde{x}_{\max} - \Delta \tilde{x}(\mathcal{T}))}{\Delta \tilde{x}_{\max}}\right]. \quad (30)$$

A positive value of $M(\mathcal{T})$ indicates that the track satisfies all acceptance criteria, whereas a negative value indicates that at least one limiting criterion is not satisfied. The sequence-level margin was defined as the largest track margin in the analysed sequence for all tracks $\mathcal{T}$ in $\mathcal{J}$:

$$M_{\text{seq}} = \max[M(\mathcal{T})]. \quad (31)$$

Candidate reflectors are first identified from the fused ultrasonic image sequence, then combined into physically admissible non-coplanar reflector pairs, linked across neighbouring axial frames and accepted only when the resulting temporal track satisfies the prescribed persistence,

score and geometric criteria. The numerical values of the final operating parameters were fixed after defining the calibration, positive and strict-background reference groups, as described in section II.H. For the sequence-level evaluation in this study, a sequence was classified as positive if at least one temporal pair track satisfied the final acceptance criteria. If multiple accepted tracks were present within the same local sequence, the sequence was counted only once and the highest-scoring accepted track was used to define the reported decision margin and representative geometric descriptors. This convention prevented multiple local tracks around the same indication from being counted as independent detections in the sequence-level performance analysis.

G. Experimental set-up and physical reference data

Experimental investigations were carried out to evaluate the proposed ultrasonic imaging and stepwise-cracking detection approach on the HIC-damaged heat-exchanger shell samples described in section II.A. The measurements were performed using plane wave imaging (PWI). During acquisition, the specimen was insonified using a sectorial plane-wave sequence covering refraction angles from -30° to $+30^\circ$ with an angular increment of 2.5° , resulting in 25 insonification angles per scan position. Although angles above $\pm 23.6^\circ$ exceeded the first critical angle for the present wedge-steel configuration, the full angular set was retained during acquisition and incoherent or mode-converted contributions were subsequently reduced by the coherence-factor weighting used in image reconstruction. The data were acquired using a 7.5 MHz, 64-element linear phased-array probe with an inter-element pitch of 0.5 mm. The active aperture length was 31.9 mm and the element elevation was 9 mm. The probe was mounted on a custom-fabricated L0-type wedge made from Rexolite, for which the longitudinal wave velocity was 2360 m/s. The wedge height was 25 mm. This height was selected to provide sufficient standoff and to reduce interference between interface-related reflections and backwall responses. Data acquisition was performed using an Eddyfi Gekko phased-array flaw detector. The probe was moved along the axial direction of the pipe using a single-axis carrier scanner. A wheel-type quadrature encoder, integrated with the Gekko system, provided position feedback with a resolution of 16 counts/mm.

For Sample S5, each axial scan consisted of 100 scan positions with a spatial increment of 1 mm. After completing one axial scan, the scanner was manually repositioned in the circumferential direction and the next axial scan was acquired. In total, 51 independent axial B-scans were collected across the circumferential direction, resulting in 5100 PWI frames. The circumferential scan lines were equally spaced to provide sufficient spatial coverage on both sides of the weld. For Sample S4, each axial scan consisted of 125 scan positions with a spatial increment of 1 mm. Ten circumferential scan positions were acquired, resulting in 10 independent axial B-scans and 1250 PWI frames. The circumferential scan lines were equally spaced to ensure coverage of the inspected section. The image reconstruction zone was defined as a 60 mm $\times$ 20 mm region, with an offset of 30 mm from the top surface of the specimen. The reconstruction grid resolution was set to 10 points/mm. The same reconstruction grid and coordinate convention were used for the subsequent coherence-factor fusion, reflector

quantification and stepwise-crack detection procedures. An illustration of the experimental setup is provided in Figure 2.

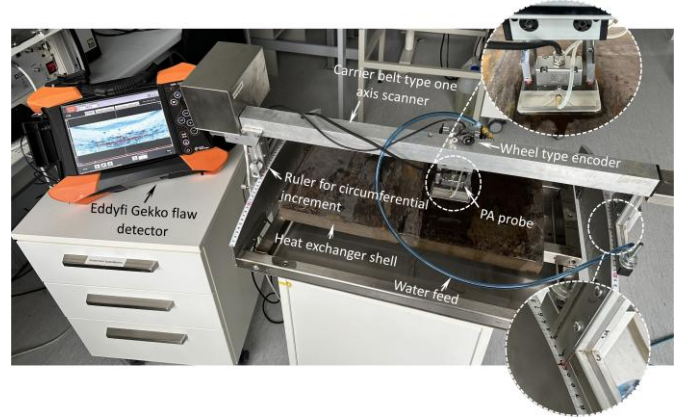


Fig. 2. Experimental set-up for assessment of HIC damage in heat exchanger shell.

The physical interpretation of the targeted stepwise-cracking signature was based on the metallographic analysis performed on sample S5. An example of the observed stepwise morphology is shown in Figure 3.

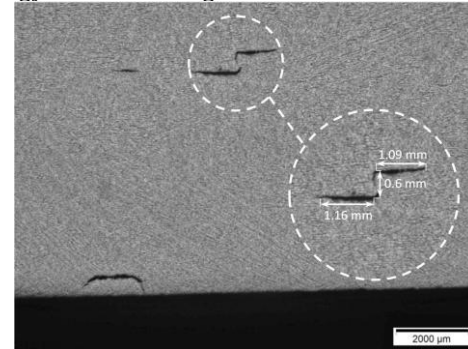


Fig. 3. Physical basis for the targeted stepwise HIC signature: a metallographic image from sample S5 showing blister interlinking and stepwise HIC morphology and geometrical interpretation of the stepwise morphology, indicating vertical separation Δz and approximate local crack extent.

The metallographic observation was used to define the physical model of the target defect morphology. In this model, stepwise HIC is interpreted as a local non-coplanar crack configuration formed by two blister-like crack segments connected through a step. Such a morphology is expected to produce an ultrasonic response containing two compact reflector components separated in depth and lateral position, rather than a single isolated reflector response. Because destructive metallographic confirmation was available only for a representative region of sample S5, the remaining real HIC indications were not treated as complete metallographic ground truth. Instead, the metallographically observed stepwise morphology was used to define the expected defect geometry and artificial calibration indications were manufactured to reproduce compact non-coplanar reflector configurations with controlled separations (see Figure 4). These calibration indications provided experimental ultrasonic references for assessing whether the imaging and detection procedure could resolve stepwise-like reflector pairs within the targeted geometrical range. The artificial calibration block therefore served as the link between the metallographic morphology and the ultrasonic detection model. The calibration-compatible

indications reproduced sub-millimetre vertical offsets comparable to the expected stepwise reflector separations and were used to examine the ultrasonic response of controlled non-coplanar reflector pairs. Based on this combined evidence, the characteristic ultrasonic signature targeted in the real HIC data was defined by four features: two local reflector components, finite vertical separation, finite lateral offset and persistence over neighbouring axial scan frames.

Additional targeted metallographic verification of the S4 indications was not undertaken. The investigated in-service shell sections are scarce specimens and destructive sectioning would permanently remove material required for further NDT investigations. More importantly, stepwise HIC is sparse, small and three-dimensional while successful confirmation requires sub-millimetre registration between the ultrasonic indication and the final metallographic section plane. The finite saw-cut kerf and the material removed during grinding and polishing can eliminate or bypass a small connecting ligament, while a nominally correct two-dimensional section may intersect only one blister-like crack segment. Consequently, the absence of a visible stepwise morphology in a single targeted section would not conclusively exclude a nearby non-coplanar defect. The reference framework used in this study therefore combines three complementary elements: representative metallographic confirmation of stepwise HIC in Sample S5, controlled ultrasonic responses from calibration-block reflector geometries of comparable scale and expert interpretation of analogous ultrasonic indications in the real-sample data.

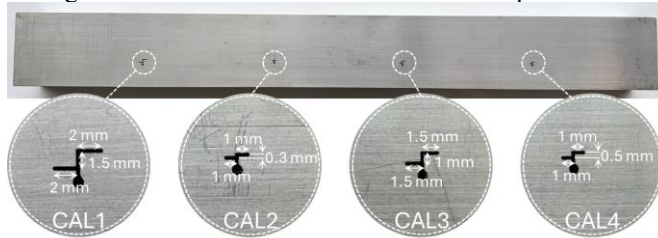


Fig. 4. Photograph of the artificial calibration block indicating the positions of CAL1–CAL4 and the intended reflector geometry

The calibration block contained four reference indications, denoted CAL1–CAL4. Their nominal geometrical properties and analysed frame intervals are summarized in Table 1.

TABLE 1

ARTIFICIAL CALIBRATION INDICATIONS USED IN THE STUDY

Calibration indication	Nominal vertical separation, Δz	Approximate axial extent	Role in study
CAL1	1.5 mm	2.0 mm	Case outside the final real-HIC operating range
CAL2	0.3 mm	1.0 mm	Primary calibration compatible case
CAL3	1.0 mm	1.5 mm	Boundary case
CAL4	0.5 mm	1.0 mm	Primary calibration compatible case

CAL2 and CAL4 were designed as compact sub-millimetre step-like reflector configurations comparable to the reflector separations targeted in the real HIC analysis. CAL1 represented a larger-offset reflector configuration outside the final operating range of the real-HIC detector, while CAL3 was

treated as a boundary case. The same imaging, reconstruction and coherence-factor-fusion procedure was applied to the calibration block and to the real HIC samples, so that both datasets were analysed within the same image-coordinate framework. Additional calibration-background and near-defect control windows were also extracted from the calibration scan, as described in section II.H.

H. Reference data groups and detector-parameter determination

The detector parameters were defined using a combination of physical prior knowledge, metallographic confirmation of one representative stepwise HIC morphology, artificial calibration indications and expert-labelled ultrasonic data. The purpose of this procedure was not to train a statistical model, but to define a fixed and deterministic operating point for the proposed physics-constrained temporal pair tracker. The physical calibration block and the nominal CAL1–CAL4 geometries were described in section II.G; the present section defines how the calibration data, real HIC indications and background controls were used for detector-parameter definition and subsequent performance evaluation.

The physical basis for the detector was obtained from a metallographically confirmed stepwise defect observed in Sample S5. The corresponding region had previously been inspected by ultrasonic TFM imaging and the metallographic section confirmed that the relevant ultrasonic indication was associated with a non-coplanar, stepwise crack morphology. This observation was used to define the expected ultrasonic signature of a stepwise HIC indication. The subsequent real HIC-positive sequences were not individually sectioned metallographically. Instead, they were labelled by an experienced NDT operator using the characteristic ultrasonic signature established from the metallographically confirmed S5 morphology and supported by the controlled calibration-block responses. The expert labelling was performed without access to the temporal-tracking scores or detector decisions. These labels were therefore treated as an expert NDT reference standard, supported by representative metallography and controlled calibration evidence, rather than as destructive ground truth for every sequence. Accordingly, the real-sample analysis should be interpreted as a fixed-parameter performance evaluation against this reference standard, not as complete diagnostic-accuracy verification of every physical stepwise crack present in the inspected shell sections. The real ultrasonic data were organized at the level of local image sequences. The HIC-positive group consisted of local sequences centred on expert-labelled HIC-like indications in Samples S4 and S5. Each real-sample sequence was extracted as a local axial window around the labelled indication, including neighbouring frames so that the temporal persistence of the reflector response could be evaluated. The expert labelling was performed blindly on the raw coherence factor fused reconstructions, without access to the temporal tracking scores.

The strict-background group consisted of local image sequences sampled away from labelled HIC indications. Background windows were selected using axial and circumferential exclusion margins relative to known positive locations, so that they represented regions without known

stepwise HIC-like indications. In the implementation used here, one circumferential scan line and twelve axial frames around known positive indications were excluded, with additional boundary margins applied during sampling. These strict-background sequences were used as NDT-defined negative controls for estimating false-positive behaviour. Since the complete samples were not destructively sectioned, strict-background windows should not be interpreted as absolute proof of crack-free material, but as regions without known or labelled stepwise HIC-like ultrasonic indications.

The calibration data were divided into artificial-defect windows and calibration-control windows. The artificial-defect group consisted of CAL1–CAL4. CAL2 and CAL4 were treated as the primary calibration-compatible cases because their nominal vertical separations, 0.3 mm and 0.5 mm, respectively, were within the sub-millimetre separation range targeted in the real HIC analysis. CAL1, with $\Delta z = 1.5$ mm, was treated as a large-offset case outside the final real-HIC operating range, while CAL3, with $\Delta z = 1.0$ mm, was treated as a boundary case. Strict calibration-background windows were selected from regions of the calibration scan away from the artificial indications. In addition, near-defect control windows were defined adjacent to selected calibration indications. These near-defect windows were not treated as strict negatives; rather, they were used to examine whether residual or tail responses from a nearby artificial reflector could persist outside the manually defined defect window. The complete composition of the reference data groups is summarized in Table 2.

The detector parameters were selected in a constrained and physically interpretable manner. First, the candidate-extraction parameters were chosen to retain compact reflector responses while suppressing isolated low-amplitude noise and small satellite peaks. Candidate selection was based on robust contrast, relative amplitude and connected-component area. These parameters determined which local image maxima were allowed to enter the pair-formation stage, but they did not by themselves define a positive HIC decision. Second, the frame-level pair-formation parameters were defined from the expected geometry of stepwise cracking. A valid candidate pair was required to satisfy both lateral and vertical separation constraints: $\Delta x_{\min} \leq \Delta x_{i,j}^{(t)} \leq \Delta x_{\max}$ and $\Delta z_{\min} \leq \Delta z_{i,j}^{(t)} \leq \Delta z_{\max}$. The lower separation limits were introduced to avoid interpreting a single broadened reflector as two independent crack steps. The upper limits restricted the detector to local stepwise morphologies rather than unrelated reflectors separated by a larger distance. The amplitude-balance threshold $B_{\min}$ was used to reject pairs in which one candidate dominated the response and the lateral decay constant λ_x was used to reduce the score of widely separated pairs within the admissible range. Third, the temporal-linking parameters were selected to reflect the expected smooth evolution of a real indication between neighbouring axial frames. The maximum frame gap $g_{\max}$ controlled the allowed temporal discontinuity, while δx_c , δz_u , δz_l and $\delta \Delta z$ controlled the allowed changes in pair centre position, upper reflector depth, lower reflector depth and vertical separation. These constraints prevented unrelated reflector pairs in neighbouring frames from being joined into a single track. Finally, the track-acceptance parameters were

chosen to define the fixed detector operating point. A track was accepted only if it satisfied the required persistence, score and median geometry criteria: $L(\mathcal{T}) \geq L_{\min}$, $S_{\max}(\mathcal{T}) \geq S_{\max,\min}$, $S_{\text{med}}(\mathcal{T}) \geq S_{\text{med},\min}$, $\Delta \tilde{z}(\mathcal{T}) \leq \Delta \tilde{z}_{\max}$, $\Delta \tilde{x}_{\min} \leq \Delta \tilde{x}(\mathcal{T}) \leq \Delta \tilde{x}_{\max}$.

TABLE 2

REFERENCE DATA GROUPS USED FOR DETECTOR-PARAMETER DEFINITION AND PERFORMANCE EVALUATION

Data group	Sample / Source	Number of sequences	Purpose in study
<i>Expert-labelled HIC-positive reference sequences</i>	S4	10	Positive real-sample reference group Aggregated sensitivity evaluation
	S5	36	
	S4 + S5	46	
<i>Strict background sequences</i>	S4	80	Negative real-sample control group Aggregated false-positive-rate and specificity evaluation
	S5	338	
	S4 + S5	418	
<i>Calibration defect windows</i>	Calibration block	4	Controlled geometric reference cases
<i>Primary calibration compatible defects</i>	Calibration block CAL2 and CAL4	2	Verification of detector response to compatible non-coplanar reflector geometry
<i>Boundary out of range defects</i>	Calibration block CAL1 and CAL3	2	Evaluation of detector selectivity relative to the targeted geometry
<i>Strict calibration background windows</i>	Calibration block	5	Calibration-block negative controls
<i>Near defect calibration control windows</i>	Calibration block	4	Examination of residual or tail responses near calibration defects
<i>All calibration control windows</i>	Calibration block	9	Calibration-control analysis; near-defect windows were not treated as strict negatives

The final operating point of the temporal pair tracker was defined from the physical stepwise-crack model, the calibration-block response and the predefined requirement to maintain high agreement with expert-labelled HIC-like reference signatures while limiting responses in strict-background regions. Once this single operating point was fixed, no adaptive refitting was performed for individual samples, scan regions or sequences; the same parameter values were applied to all real HIC-positive sequences, strict-background sequences and calibration windows reported in Section III. The selected final operating parameters are listed in Table 3. The frame-level pair-formation limits were intentionally slightly wider than the final track-level acceptance limits. This allowed

potentially relevant reflector pairs to enter the temporal-linking stage.

TABLE 3

FINAL PARAMETERS OF THE PROPOSED TEMPORAL PAIR-TRACKING DETECTOR

Parameter group	Parameter	Value	Unit
Candidate extraction	Minimum robust contrast, $R_{\min}$	5.0	–
	Minimum relative amplitude	0.01	–
	Maximum candidates per frame	40	Candidates
Pair formation	$\Delta z_{\min} - \Delta z_{\max}$	0.2 – 1.0	mm
	$\Delta x_{\min} - \Delta x_{\max}$	0.5 – 1.5	mm
	$B_{\min}$	0.20	–
	λ_x	2.5	mm
Pair-to-track gating	Minimum pair score	0.05	–
Temporal linking	Maximum frame gap, $g_{\max}$	1	frame
	Maximum pair-centre shift, δx_c	2.0	mm
	Maximum upper/lower depth shift, $\delta z_u, \delta z_l$	0.5	mm
	Maximum Δz change, $\delta \Delta z$	0.4	mm
Track acceptance	$L_{\min}$	3	frames
	$S_{\max, \min}$	0.20	–
	$S_{\text{med}, \min}$	0.15	–
	$\Delta \tilde{z}_{\max}$	0.8	mm
	$\Delta \tilde{x}_{\min} - \Delta \tilde{x}_{\max}$	0.7 – 1.5	mm

I. Performance and robustness evaluation

The final detector was evaluated after fixing the reference data groups and operating parameters described in section II.H. The evaluation was performed at the level of local image sequences or calibration windows. The real HIC-positive sequences, strict-background sequences, calibration artificial-defect windows and calibration-control windows used for this evaluation are defined in Table 2, while the fixed detector parameters are listed in Table 3. No sequence-specific parameter adjustment was performed during the evaluation.

For each analysed sequence $\mathcal{J}$, the detector returned a binary sequence-level decision $D(\mathcal{J})$, where $D(\mathcal{J}) = 1$ indicates that at least one temporal reflector-pair track satisfied the final acceptance criteria and $D(\mathcal{J}) = 0$ indicates that no accepted track was found. For the real HIC-positive group, a detected sequence was counted as a true positive (TP). For the strict-background group, a detected sequence was counted as a false positive (FP). The primary performance metrics were therefore sequence-level sensitivity, false-positive rate and specificity. The sensitivity was estimated as:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \quad (32)$$

where TP is the number of detected HIC-positive sequences and FN is the number of missed HIC-positive sequences (false negatives). The false-positive rate was calculated as:

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}, \quad (33)$$

where TN is the number of strict-background sequences rejected by the detector. Specificity was calculated as:

$$\text{Specificity} = 1 - \text{FPR}, \quad (34)$$

The primary performance values were calculated for the aggregated S4 + S5 data groups. Sample-wise values were also calculated to verify consistency between the two real samples, but the aggregated values were used as the main performance

estimates because they better represent the full available real-sample reference set. Sample-wise values were interpreted descriptively because the number of expert-labelled positive sequences in each individual shell section was too small to support precise sample-specific sensitivity estimates. Here, TP and FN are defined relative to the expert NDT reference labels. Therefore, the reported sensitivity quantifies agreement with the expert reference standard and should not be interpreted as a metallographically verified probability of detecting every physical stepwise crack present in the inspected material.

To quantify statistical uncertainty in the sequence-level rates, Wilson score confidence intervals were calculated for sensitivity and false-positive rate. For a binomial proportion $\hat{p} = k/n$, where k is the number of events and n is the number of evaluated sequences, the 95% Wilson interval was calculated as:

$$CI_{\text{Wilson}} = \frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{(\hat{p}(1-\hat{p}))}{n} + \frac{z^2}{4n^2}}}{\left(1 + \frac{z^2}{n}\right)} \quad (35)$$

where z is the critical value from the standard normal distribution. For a 95% confidence interval, this value is 1.96. The Wilson interval was used because it provides more stable uncertainty estimates than the normal approximation when the number of events is limited or when the measured rate is close to zero or one. The aggregated sensitivity, false-positive rate, specificity and corresponding uncertainty intervals were calculated using the procedure described above and are reported in the section III.C.

Because multiple local sequences may originate from the same sample and circumferential scan line, the independence assumption of a simple binomial model may be optimistic. Therefore, a cluster-bootstrap analysis was also performed. Clusters were defined by sample and circumferential scan index. In each bootstrap iteration, clusters were sampled with replacement and the corresponding sequence-level sensitivity or false-positive rate was recalculated. The 2.5th and 97.5th percentiles of the bootstrap distribution were used as the empirical 95% cluster-bootstrap interval. This analysis was used to assess whether the measured performance remained stable when correlation between sequences from the same scan region was considered.

A leave-one-cluster-out analysis was additionally performed to examine whether the measured performance was dominated by any individual circumferential scan region. In this analysis, one sample/circumferential-scan cluster was removed at a time and the aggregated sensitivity or false-positive rate was recalculated from the remaining data. The range of leave-one-cluster-out values was used as an influence measure. The leave-one-cluster-out results are reported together with the robustness analysis in the section III.F.

The decision-margin formulation introduced in section II.F was used to provide an interpretable measure of the distance from the final detection boundary. For each sequence, the best-track sequence margin M_{seq} was calculated as the largest margin among all candidate tracks in the sequence. A positive value of M_{seq} indicates that the best track satisfies all final acceptance criteria, while a negative value indicates that at least one criterion remains unsatisfied. The distribution of M_{seq} was compared between real HIC-positive and strict-background

sequences. This analysis was used to determine whether the final detector separated the two reference groups by a transparent multi-criterion decision boundary rather than by an isolated amplitude threshold. The decision-margin distributions for real HIC-positive and strict-background sequences were generated after applying the final fixed detector and are presented in the section III.E.

To evaluate the contribution of the individual detector components, an ablation study was performed. The same candidate tracks were evaluated using five rule variants: score-only, persistence-only, score plus persistence, score plus geometry and the final score plus persistence plus geometry rule. The score-only variant used the maximum and median pair-score criteria without the persistence and final geometry criteria. The persistence-only variant used only the minimum track persistence criterion. The score-plus-persistence variant combined score and persistence criteria, while the score-plus-geometry variant combined score and final track-level geometry without the temporal persistence criterion. The full detector used all final criteria. For each ablation variant, sensitivity, false-positive rate and specificity were recalculated using the same real HIC-positive and strict-background sequence sets. This analysis was used to determine whether the final detector performance resulted from the combined use of score, persistence and geometry rather than from a single threshold. The ablation analysis was performed using the same reference data groups and fixed candidate tracks. The resulting sensitivity, false-positive rate and specificity values for each rule variant are reported in the section III.E.

A local parameter-stability sweep was performed around the final operating point to assess whether the selected parameter set represented a stable region of acceptable performance. The sweep varied the main final track-acceptance parameters, including minimum track persistence, maximum pair-score threshold, median pair-score threshold, maximum median vertical separation and median lateral-separation limits. For each parameter combination, the detector was re-evaluated on the same real HIC-positive and strict-background groups. Candidate operating points were identified as those maintaining sensitivity above a predefined minimum level while keeping the false-positive rate below a predefined maximum level. This analysis was used only as a robustness check of the final operating point; the final parameter values used for the main evaluation remained those listed in Table 3.

A separate detector-level fusion comparison was also performed to examine whether the selection of VCF×CCF was specific to the fixed final operating point. For this analysis, all six pairwise coherence-factor fusion variants were evaluated using the same real HIC-positive and strict-background sequence locations. Two complementary comparisons were made. First, the final VCF×CCF operating point was applied unchanged to all fusion variants. Second, for each fusion variant, the final track-acceptance parameters were swept independently over the same local parameter grid and the best fusion-specific operating point was selected by maximizing the Youden index, $J = \text{sensitivity} + \text{specificity} - 1$. The candidate-extraction parameters, pair-formation rules and temporal-linking parameters were kept fixed in this comparison, so that the physical detector structure remained identical across fusion

methods. Although the candidate-extraction stage is influenced by image statistics, these upstream parameters were not retuned separately for each fusion method because doing so would create fusion-specific detector front ends and confound the comparison of image representations with additional hyperparameter selection. Instead, the same adaptive candidate-extraction procedure, based on per-frame robust contrast and relative-amplitude normalization, was applied to all fusion methods, while only the final track-acceptance operating point was optimized independently. This analysis was used to determine whether each fused representation could support a favourable sensitivity–false-positive-rate operating regime under its own optimized final track-acceptance thresholds.

To further examine whether the selected operating point was dependent on using both real samples together, a cross-sample holdout stress test was performed. Because sample S5 contained the larger reference set, with 36 expert-labelled positive sequences and 338 strict-background sequences, the final track-acceptance parameters were selected using only S5. The selected rule was then applied without modification to the holdout S4 sequences. During this analysis, the detector structure, candidate-extraction parameters, pair-formation rules and temporal-linking parameters were kept unchanged, only the final track-acceptance parameters were selected from the same local parameter grid used in the stability sweep. The S5-selected operating point was chosen by maximizing the Youden index among parameter combinations satisfying sensitivity $\geq 80\%$ and false-positive rate $\leq 10\%$ on S5. Because S4 contained only 10 positive reference sequences, this analysis was interpreted as a cross-sample stress test rather than a statistically powered external validation.

The calibration block was evaluated separately to verify the geometric consistency of the detector response. The same fixed detector parameters were applied to the CAL1–CAL4 artificial-defect windows and to the calibration-control windows defined in section II.H. For the calibration-compatible artificial defects, the median detected vertical separation $\Delta\tilde{z}(\mathcal{T})$ was compared with the nominal calibration separation Δz_{nom} . The absolute vertical-separation error was calculated as:

$$e\Delta z = |\Delta\tilde{z}(\mathcal{T}) - \Delta z_{\text{nom}}|. \quad (36)$$

The calibration analysis was used to verify that the final detector could respond to controlled compact non-coplanar reflector geometries within the targeted stepwise range and to examine whether strict calibration-background and near-defect control windows produced accepted tracks. The calibration compatibility analysis was performed by applying the final fixed detector to the CAL1–CAL4 artificial-defect windows and calibration-control windows. The resulting detection decisions and Δz measurement errors are reported in the section III.B.

Together, these analyses evaluated the detector from complementary perspectives: sequence-level sensitivity to expert-labelled HIC-like indications, false-positive behaviour in strict-background regions, uncertainty due to limited and clustered data, influence of individual scan regions, contribution of score/persistence/geometry criteria, stability of the selected operating point and geometric consistency on controlled calibration indications.

III. RESULTS AND DISCUSSION

A. Coherence factor fusion results

In this section, we present the results of pixel-wise multiplication between all pairwise combinations of coherence factors, as outlined in section II.C. To quantitatively evaluate the enhancement in image contrast and resolution, we utilize the CNR and API metrics introduced in section II.E. For demonstration purposes, a single PWI frame was selected from the axial scan of the experimental sample described in section II.A, acquired using the experimental setup detailed in section II.G. This particular frame contains a stepwise crack. For the original PWI image, the individual VCF, CCF, SCF and PCF coherence factors were applied, and the corresponding CNR and API values were estimated, as illustrated in Fig. 5. Figure 6 shows the results obtained from pixel-wise fusion of these coherence factors in six pairwise combinations: VCF×CCF, VCF×SCF, VCF×PCF, SCF×CCF, SCF×PCF and CCF×PCF. The quantitative image-quality comparison for the original PWI image, the individual coherence-factor-weighted images and the pairwise fusion strategies is summarized in Table 4.

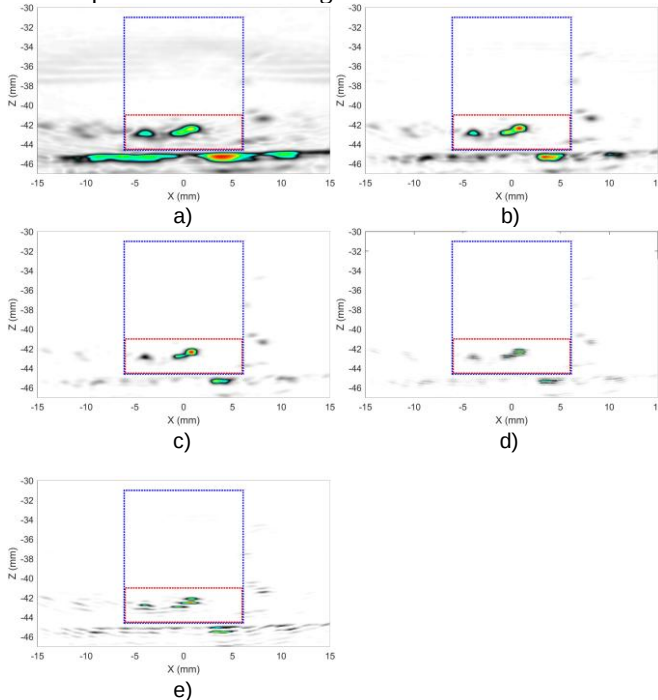


Fig. 5. Individual coherence factors: a) the original PWI image, b) VCF, c) CCF, d) SCF and e) PCF coherence factor corrections. All images are displayed using colour scales normalized to the maximum magnitude in each image. The red square denotes the signal region and the blue square denotes the noise region used for CNR and API estimation.

As summarized in Table 4, by applying individual coherence factor, the CNR of the original PWI image can be improved up to 8 dB, from 16.8 dB in the original PWI image to 24.8 dB in CCF reconstruction. Notably, among the individual coherence factors the best results are obtained with CCF and to some extent with VCF factors. In the PCF and SCF images, the indication corresponding to the stepwise crack appears spatially extended, with multiple reflection lobes distributed over several axial depths. This diffused appearance is indicative of reduced axial resolution and suggests the reduction in the clarity of the true reflector geometry. In contrast, the CCF and

VCF images exhibit a more compact and localized indication of the same defect. This suggests enhanced spatial resolution and contrast, which can be attributed to CCF's robustness in suppressing off-axis and phase-incoherent contributions while reinforcing signals that exhibit circular phase consistency.

The pairwise fusion results in Table 4 show that the CNR was further improved by up to 12.1 dB compared with conventional PWI/TFM. The highest CNR and lowest API were obtained using SCF×CCF. However, CNR and API do not fully describe whether the fused image is suitable for the proposed temporal pair-tracking detector, because the detector operates on compact local reflector candidates and physically admissible reflector pairs.

TABLE 4

IMAGE-QUALITY COMPARISON OF INDIVIDUAL COHERENCE FACTORS AND COHERENCE-FACTOR FUSION STRATEGIES

Weighting method	CNR (dB)	CNR gain vs. TFM (dB)	API	API reduction vs. TFM (%)
None	16.8	-	7.77	-
VCF	21.7	4.9	2.89	62.8
CCF	24.8	8.0	0.78	90.0
SCF	22.6	5.8	0.37	95.2
PCF	21.2	4.4	0.74	90.5
VCF×CCF	27.7	10.9	0.47	94
VCF×SCF	25.6	8.8	0.19	97.6
VCF×PCF	24.8	8.0	0.32	95.9
SCF×CCF	28.9	12.1	0.15	98.1
SCF×PCF	27.2	10.4	0.18	97.7
CCF×PCF	27.2	10.4	0.27	96.5

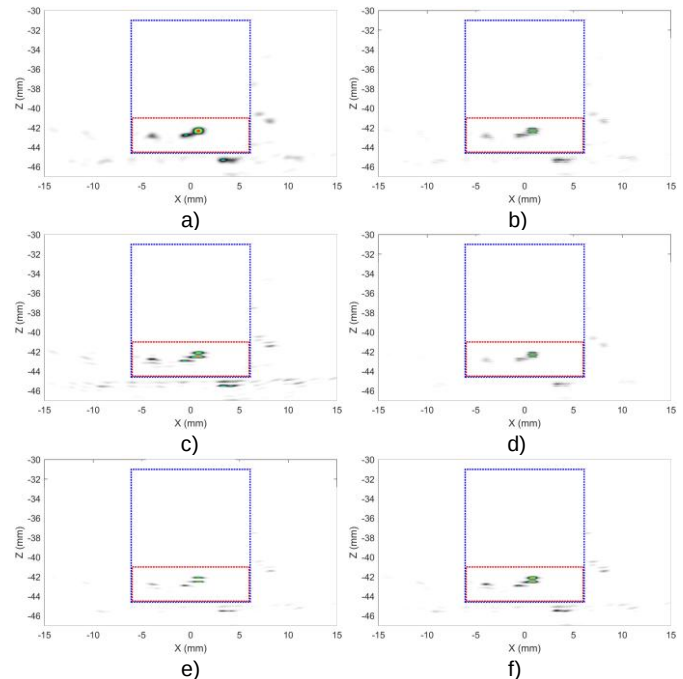


Fig. 6. The original PWI image corrected with pairwise fusion of the coherence factors: a) VCF×CCF, b) VCF×SCF, c) VCF×PCF, d) SCF×CCF, e) SCF×PCF and f) CCF×PCF. All images are displayed using colour scales normalized to the maximum magnitude in each image. The red square denotes the signal region and the blue square denotes the noise region used for CNR and API estimation.

To quantify reflector morphology more objectively, connected-component metrics were calculated in the same signal region used for the CNR and API analysis. The main reflector pair was not selected solely by component size, but

was selected from component pairs satisfying the same frame-level geometric admissibility limits as the detector. The results are summarized in Table 5. VCF×CCF produced the highest main-pair compactness, 0.599, whereas SCF×CCF, despite its superior CNR and API values, produced a substantially lower compactness value of 0.251. The selected VCF×CCF component pair had centroid separations of $\Delta x = 1.09$ mm and $\Delta z = 0.48$ mm, which lie within the detector-compatible frame-level geometry. VCF×CCF and SCF×CCF both produced three connected components and comparable satellite area fractions outside the selected main pair, 24.0% and 23.5%, respectively, but the much higher compactness of VCF×CCF indicates a more regular and localized main reflector-pair morphology. Other PCF- or SCF-containing combinations produced four to five connected components and satellite area fractions between 25.0% and 47.8%, indicating greater fragmentation of the thresholded response. Thus, among the tested fusion strategies, VCF×CCF provided the most favourable compromise between contrast improvement, artefact suppression and preservation of compact detector-compatible reflector-pair morphology. It increased the CNR of the original PWI image by 10.9 dB and reduced the API from 7.77 to 0.47. Compared with the use of individual coherence factors, VCF×CCF increased the CNR by 2.9–6.5 dB. VCF×CCF was therefore selected as the primary fused representation for the subsequent temporal pair-tracking analysis. Because CNR, API and morphology metrics remain image-level descriptors, the choice of fusion strategy was further examined using the end-to-end detector-level comparison reported in Section III.D.

TABLE 5
CONNECTED COMPONENT MORPHOLOGY OF PAIRWISE FUSION
RESPONSES IN THE REPRESENTATIVE SIGNAL REGION

Fusion method	Connected components	Thresholded area (mm ²)	Main pair compactness	Satellite area outside main pair (%)
VCF×CCF	3	0.50	0.599	24.0
VCF×SCF	4	0.26	0.196	30.8
VCF×PCF	4	0.32	0.342	31.3
SCF×CCF	3	0.17	0.251	23.5
SCF×PCF	5	0.23	0.374	47.8
CCF×PCF	4	0.20	0.419	25.0

The morphology metrics were calculated in the same signal region used for Fig. 6 and Table 4. The −6 dB threshold was used to quantify the high-intensity reflector core, consistent with the API definition. The main-pair compactness was calculated from the detector-compatible component pair with the largest combined thresholded area among pairs satisfying the frame-level centroid-separation limits $\Delta x = 0.5$ – 1.5 mm and $\Delta z = 0.2$ – 1.0 mm. The satellite area fraction is the thresholded area outside this selected pair divided by the total thresholded area.

B. Calibration-block response of the final detector

The artificial calibration block was first used to examine whether the final detector responded to controlled non-coplanar reflector geometries within the targeted stepwise range. The VCF×CCF fused reconstructions of the four calibration indications are shown in Figure 7. These images illustrate the different geometrical conditions represented by CAL1–CAL4

and provide a direct visual comparison between the calibration-compatible cases, the boundary case and the large-offset case.

The final temporal pair-tracking detector was then applied to the CAL1–CAL4 windows without changing the operating parameters listed in Table 3. The resulting detection decisions and measured track properties are summarized in Table 6. The two calibration-compatible indications, CAL2 and CAL4, were both detected by the final operating point. For CAL2, which had a nominal vertical separation of 0.3 mm, the detector measured a median track separation of 0.4 mm. For CAL4, which had a nominal vertical separation of 0.5 mm, the measured median track separation was 0.6 mm. In both cases, the absolute Δz error was 0.1 mm which corresponds to a single pixel value.

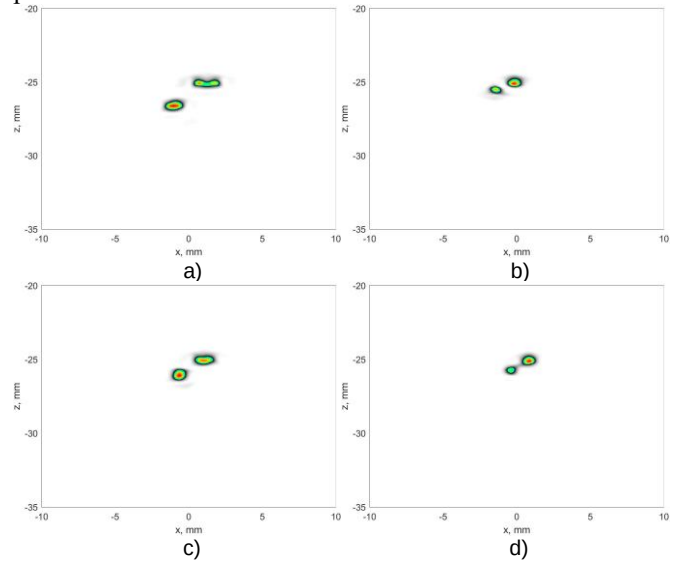


Fig. 7. VCF×CCF fused reconstructions of the artificial calibration indications: a) CAL1, large-offset case with nominal $\Delta z = 1.5$ mm; b) CAL2, calibration-compatible compact step case with nominal $\Delta z = 0.3$ mm; c) CAL3, boundary case with nominal $\Delta z = 1.0$ mm; d) CAL4, calibration-compatible compact step case with nominal $\Delta z = 0.5$ mm.

TABLE 6
RESPONSE OF THE FINAL DETECTOR TO THE ARTIFICIAL CALIBRATION INDICATIONS

Case	Detector decision	Track persistence	Median pair score	Detected Median Δz	Detected median Δx
CAL1	Rejected	0 frames	-	-	-
CAL2	Detected	10 frames	0.872	0.4 mm	1.3 mm
CAL3	Rejected	1 frame	0.307	0.2 mm	0.9 mm
CAL4	Detected	11 frames	1.210	0.6 mm	1.2 mm

The detected calibration-compatible cases also produced long and stable tracks. CAL2 was tracked over 10 PWI frames, while CAL4 was tracked over 11 frames. This behaviour is consistent with the expected response of compact non-coplanar reflector pairs that remain visible over neighbouring axial scan positions. In contrast, CAL1 was rejected because its nominal vertical separation was outside the final real-HIC operating range. CAL3 was also rejected by the final detector and behaved as a boundary case: although a weak one-frame candidate track was formed, it did not satisfy the required temporal persistence and final track-level acceptance criteria.

The calibration-control windows were analysed to distinguish strict background behaviour from near-defect

residual responses. All five strict calibration-background windows were rejected by the final detector. Among the four near-defect control windows, one window adjacent to CAL2 was detected. This near-defect response was not treated as a strict-background false positive because the window was located next to an artificial reflector and may contain residual or tail response from the nearby calibration indication. This observation supports the use of exclusion margins around known indications when strict-background windows are defined in the real HIC samples. Because such adjacent-reflector responses could potentially generate spatially redundant accepted tracks, the number of accepted tracks per sequence was quantified for the real HIC-positive and strict-background results, as discussed in Section III.C.

These calibration results show that the final detector responds to compact non-coplanar reflector geometries within the targeted sub-millimetre range while rejecting the larger-offset and boundary calibration cases under the fixed real-HIC operating point. The calibration block should therefore be interpreted as a controlled ultrasonic reference for the proposed stepwise-reflector model, rather than as a metallurgical replica of hydrogen-induced cracking.

C. Sequence-level detection performance on real HIC and strict-background data

After confirming the response of the detector on the calibration block, the final fixed operating point was applied to the real HIC-positive and strict-background sequences defined in Section II.H. The purpose of this analysis was to evaluate whether the proposed temporal pair-tracking detector could identify expert-labelled stepwise-HIC-like ultrasonic indications while maintaining a low response rate in regions without known HIC-like indications. Accordingly, the term sensitivity in this section denotes sensitivity relative to the expert-labelled reference sequences. It represents agreement with the available expert NDT reference standard and not a complete metallographic probability-of-detection estimate for all possible stepwise-HIC morphologies. The detector parameters were not adjusted separately for samples S4, S5 or individual sequences. The example of the detected HIC damage on the S5 sample is presented in Figure 8.

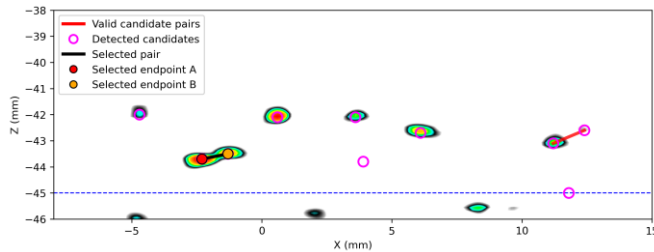


Fig. 8. Example accepted stepwise-HIC-like indication. The fused image is shown for a representative frame of the accepted temporal track. Magenta markers indicate detected candidate peaks participating in valid pairs, red line indicates candidate peak pairs in the displayed frame and black line indicates the final selected pair.

The sequence-level performance results are summarized in Table 7. For the combined S4 and S5 real HIC-positive group, the detector identified 41 of 46 labelled HIC-positive sequences, corresponding to a sensitivity of 89.1%. For the combined strict-background group, 27 of 418 background

sequences were classified as positive, corresponding to a false-positive rate of 6.5% and a specificity of 93.5%. The aggregated S4 + S5 values are used as the primary performance estimates because they include the full available real-sample reference set.

The Wilson confidence intervals quantify the uncertainty associated with the finite number of evaluated sequences. For the real HIC-positive group, the 95% Wilson interval for sensitivity was 77.0–95.3%. For the strict-background group, the 95% Wilson interval for the false-positive rate was 4.5–9.2%. Thus, even the upper confidence bound for the false-positive rate remained below 10%. The corresponding specificity was 93.5%, with a Wilson 95% interval of 90.8–95.5%. The width of the sensitivity interval reflects the limited number of positive reference sequences available for this rare defect morphology. At the observed sensitivity level, the available 46 positive sequences provide an approximate Wilson half-width of 9.2 percentage points, which is consistent with an initial fixed-parameter reference-standard evaluation. A Wilson-based planning calculation indicates that approximately 150 independent positive indications would be required to obtain a 95% interval with an approximate ± 5 percentage-point half-width. For a stricter qualification criterion requiring the lower Wilson bound to exceed 85%, approximately 270 positive indications would be required. In contrast, the false-positive rate was estimated more precisely because 418 strict-background sequences were available. Its Wilson 95% interval was 4.5–9.2%, so even the upper confidence bound remained below 10%. Therefore, the present result should be interpreted as a fixed-parameter reference-standard evaluation rather than as a qualification-level probability-of-detection study.

Because multiple local sequences could originate from the same sample and circumferential scan region, a cluster-bootstrap analysis was also performed. In this analysis, clusters were defined by sample and circumferential scan index and clusters were resampled with replacement. The resulting bootstrap intervals were consistent with the Wilson intervals. The cluster-bootstrap interval was 79.5–97.7% for sensitivity and 4.2–8.8% for the false-positive rate. This agreement indicates that the measured performance was not strongly affected by the correlation between local sequences from the same scan region.

TABLE 7

SEQUENCE-LEVEL PERFORMANCE OF THE FINAL DETECTOR AGAINST EXPERT-LABELLED HIC-POSITIVE AND STRICT-BACKGROUND REFERENCE SEQUENCES

Data group	Number of sequences	Performance rate	Wilson 95% CI	Cluster bootstrap 95% interval
Expert labelled HIC-positive sequences	46	Sensitivity: 89.1%	77.0% – 95.3%	79.5% – 97.7%
Strict-background sequences	418	False-positive rate: 6.5% Specificity: 93.5 %	4.5% – 9.2% 90.8% – 95.5%	4.2% – 8.8% 91.2% – 95.8%

Sample-wise performance was also examined descriptively. In sample S4, the detector identified 9 of 10 real HIC-positive

reference sequences and produced 4 positive calls among 80 strict-background sequences. In sample S5, the detector identified 32 of 36 real HIC-positive reference sequences and produced 23 positive calls among 338 strict-background sequences. This corresponds to sensitivity values of 90.0% and 88.9% and false-positive rates of 5.0% and 6.8% for S4 and S5, respectively. These sample-wise point estimates were similar, but they are reported only as descriptive consistency checks because the number of positive reference sequences in each individual sample, particularly S4, was limited. The aggregated S4 + S5 result is therefore used as the primary performance estimate.

In addition to the sequence-level detection decision, the accepted temporal tracks provide apparent ultrasonic geometrical descriptors of the detected stepwise-HIC-like indications. For each detected real HIC-positive sequence, the highest-scoring accepted track was used to estimate the median vertical reflector separation Δz , median lateral offset Δx , apparent inclination angle $\theta = \tan^{-1}(\Delta z/\Delta x)$. The results are presented in Table 8. These quantities should be interpreted as image-derived reflector-pair descriptors rather than direct metallographic crack dimensions, because destructive confirmation was not available for every indication.

TABLE 8
APPARENT ULTRASONIC GEOMETRICAL DESCRIPTORS OF THE 41
DETECTED REAL HIC-POSITIVE SEQUENCES.

Descriptor	Median	IQR	Min-max
<i>Median vertical separation Δz, mm</i>	0.4	0.3-0.5	0.2-0.8
<i>Median lateral offset Δx, mm</i>	1.0	0.9-1.2	0.7-1.5
<i>Apparent inclination angle θ, °</i>	18.4	16.7-24.0	9.5-45.0

The number of accepted tracks per sequence was also analysed to assess spatial redundancy of the detector output. Across the 46 expert-labelled positive reference sequences, five sequences had no accepted track, 38 had one accepted track and three had two accepted tracks; no positive sequence contained three or more accepted tracks. Thus, among the 41 detected positive sequences, 38 sequences (92.7%) were represented by a single accepted track and three sequences (7.3%) by two accepted tracks. In the three multi-track positive sequences, the closest distance between median track centres ranged from 0.35 mm to 1.08 mm, with a median of 1.03 mm, indicating that multiple accepted tracks were rare and remained localized within the same sequence-level indication. Among the 27 strict-background sequences classified as positive, 26 sequences (96.3%) contained one accepted track and one sequence (3.7%) contained two accepted tracks. Therefore, the reported sequence-level sensitivity and false-positive rate were not inflated by multiple-track counting, because each local sequence contributed only one binary decision. However, this sequence-level analysis does not by itself resolve the inverse indication-counting problem, where two physically distinct but closely spaced HIC clusters could be represented by a single accepted reflector-pair track if their responses fall within the same local pairing and temporal-linking tolerances.

The present detector should therefore be interpreted as a sequence-level decision tool rather than as a complete

indication-counting algorithm. For future indication-level mapping over larger scan areas, accepted tracks should be post-processed using spatial non-maximum suppression or clustering of track centres. Such a de-duplication stage would merge redundant tracks located within the same local neighbourhood, while retaining spatially separated accepted tracks as distinct candidate indications when their centre separation exceeds a calibrated reporting radius. The reporting radius and separation rule should be validated using calibration specimens containing multiple known non-coplanar reflector pairs at controlled lateral and axial separations.

D. Detector-level comparison of fusion strategies

The CNR, API and connected-component morphology results in Section III.A quantify image-level contrast, resolution and reflector compactness for a representative frame. However, these metrics do not directly measure whether a fused image is suitable for the proposed temporal pair-tracking detector. A fusion method may improve apparent image quality while also altering the local morphology of reflector responses, fragmenting indications, or producing artefact structures that can form physically admissible but false candidate pairs. Therefore, the fusion strategies were also compared at the detector level.

The detector-level comparison was performed using the same real HIC-positive and strict-background sequence locations for all fusion methods. Two complementary criteria were used, as summarized in Table 9. The fixed-rule comparison applied the final VCF×CCF operating point unchanged to every fusion method. The optimized comparison then swept the final track-acceptance thresholds independently for each fusion method and selected the fusion-specific operating point that maximized the Youden index. The candidate-extraction parameters, pair-formation rules and temporal-linking parameters were kept fixed, so that all fusion methods were compared using the same physics-constrained detector structure. Consequently, the optimized values in Table 9 should be interpreted as fusion-specific final-decision operating points within a common temporal pair-tracking architecture, rather than as exhaustive independent optimization of all front-end and tracking parameters for each fusion method.

The results show that VCF×CCF remained the most favourable fusion strategy even when each fusion method was allowed to use its own optimized final track-acceptance thresholds. With the fixed final rule, VCF×CCF achieved 89.1% sensitivity and 6.5% false-positive rate. Under fusion-specific optimization, VCF×CCF again achieved the highest maximum Youden index, $J = 0.83$, with 89.1% sensitivity and 6.5% false-positive rate. The closest alternative, VCF×SCF, reached $J = 0.70$ with 84.8% sensitivity and 14.8% false-positive rate. Although SCF×CCF produced the highest CNR and lowest API in the image-quality comparison, its optimized operating point required a substantially higher false-positive rate of 25.6% to reach 87.0% sensitivity. These results indicate that local image-quality metrics alone are insufficient for selecting the fused representation for automated stepwise-HIC detection. The fused image must also preserve compact, temporally persistent reflector-pair morphology compatible with the temporal pair-tracking detector. As shown in Table 9,

when the false-positive rate was constrained to be no greater than 10%, VCF×CCF also provided the highest achievable sensitivity, 91.3%, whereas the next-best fusion strategy remained below 75% sensitivity under the same false-positive-rate constraint.

TABLE 9

DETECTOR-LEVEL COMPARISON OF PAIRWISE COHERENCE-FACTOR FUSION METHODS USING FIXED-RULE AND FUSION-SPECIFIC OPTIMIZED FINAL TRACK ACCEPTANCE PARAMETERS*

Fusion method	Fixed final rule Sens/FPR (%)	Optimized final track parameters	Optimized Sens/FPR at max J (%)	Maximum Youden index, J
VCF×CCF	89.1/6.5	3 / 0.20 / 0.15 / 0.8 / 0.7–1.5	89.1/6.5	0.83
VCF×SCF	69.6/14.8	2 / 0.20 / 0.15 / 0.6 / 0.7–1	84.8/14.8	0.70
SCF×CCF	69.6/11.5	2 / 0.20 / 0.15 / 0.6 / 0.5–1.5	87.0/25.6	0.61
VCF×PCF	82.6/24.9	3 / 0.15 / 0.15 / 0.6 / 0.7–1.3	76.1/17.5	0.59
CCF×PCF	69.6/20.1	2 / 0.15 / 0.10 / 0.6 / 0.7–1.3	80.4/21.5	0.59
SCF×PCF	67.4/23.2	3 / 0.25 / 0.20 / 0.6 / 0.5–1.5	67.4/17.9	0.49

*The optimized final-track parameter column lists $L_{\min}$ / $S_{\max,\min}$ / $S_{\text{med},\min}$ / $\Delta\tilde{z}_{\max}$ / $\Delta\tilde{x}_{\min}$ – $\Delta\tilde{x}_{\max}$. Units are frames, dimensionless score, dimensionless score, mm and mm, respectively. For each fusion method, the optimized operating point was selected independently by maximizing J over the same final track-acceptance parameter grid.

E. Decision-margin and ablation analysis

The sequence-level decision-margin analysis was used to examine whether the final detector separated real HIC-positive and strict-background sequences by a physically interpretable decision boundary. As defined in section II.F, the decision margin represents the minimum normalized distance of the best track from the final acceptance criteria. Therefore, positive margin values indicate accepted tracks, whereas negative margin values indicate rejected sequences for which at least one final criterion was not satisfied. The resulting decision-margin distributions are shown in Figure 9(a).

All 46 real HIC-positive sequences produced at least one candidate temporal pair track. In contrast, only 245 of 418 strict-background sequences produced any candidate track, while the remaining 173 strict-background sequences were rejected before final track scoring. These sequences were not used in the Figure 9(a). Most real HIC-positive sequences were located on the positive side of the decision boundary, while most strict-background sequences were located on the negative side. This demonstrates that the final detector did not rely on a single amplitude threshold, but instead combined temporal persistence, reflector-pair score and physically admissible pair geometry into one interpretable decision variable. The decision-margin distribution in Figure 9(a) also provides a practical confidence interpretation of the detector output. A large positive margin indicates that the accepted track satisfies all criteria with reserve, whereas a margin close to zero indicates a borderline indication. Conversely, a negative

margin identifies a sequence in which the best available track did not satisfy at least one requirement, such as persistence, score, or final median geometry. This allows accepted and rejected sequences to be ranked according to their proximity to the final decision boundary.

Figure 9(b) shows how sequence-level sensitivity and false-positive rate vary as the decision-margin threshold is swept. The selected operating threshold, $M_{\text{seq}} = 0$, corresponds to accepting any track that satisfies all final acceptance criteria and coincides with the fixed operating point reported in Table 7. Increasing the threshold above zero further reduced the false-positive rate but at a progressively larger cost in sensitivity, confirming that $M_{\text{seq}} = 0$ provided a balanced trade-off consistent with the parameter-stability analysis in Section III.F.

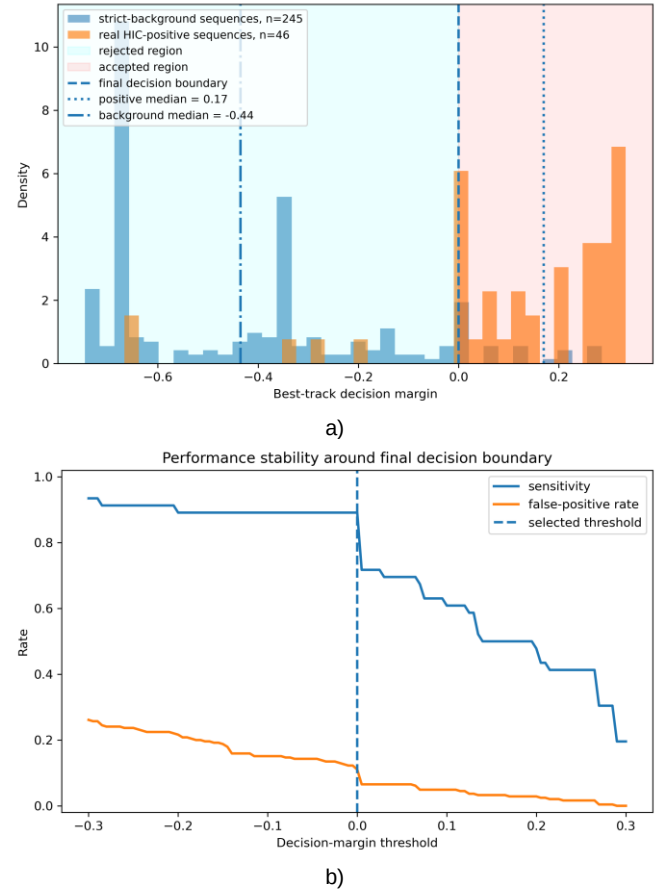


Fig. 9. Sequence-level decision-margin analysis. a) distribution of the sequence-level decision margin M_{seq} for real HIC-positive and strict-background sequences (the vertical line at $M_{\text{seq}} = 0$ indicates the final decision boundary), b) sensitivity and false-positive rate obtained by varying the decision-margin threshold (the dashed vertical line indicates the selected threshold $M_{\text{seq}} = 0$ used for the final detector decision).

To evaluate the contribution of the individual detector components, an ablation study was performed using the same real HIC-positive and strict-background sequence groups. The results are summarized in Table 10.

The score-only rule detected 43 of 46 positive sequences, corresponding to a sensitivity of 93.5%, but also produced 87 positive calls among 418 strict-background sequences, corresponding to a false-positive rate of 20.8%. This shows that pair-score information alone was sensitive to HIC-like indications, but insufficiently specific. The persistence-only

rule also detected 43 of 46 positive sequences, but produced the highest false-positive rate, 30.4%. This indicates that temporal persistence alone is not sufficient, because some background reflector configurations also persist over adjacent frames. Combining score and persistence reduced the false-positive rate to 9.81% while retaining 91.3% sensitivity. The score plus geometry rule achieved the same sensitivity of 91.3%, but with a higher false-positive rate of 15.1%. Therefore, final track-level geometry was useful, but geometry without temporal persistence did not provide the same level of background rejection. The full detector, combining score, persistence and geometry, produced the lowest false-positive rate of 6.5% and the highest specificity of 93.5%, while maintaining 89.1% sensitivity. Compared with the score-only rule, the full detector reduced the false-positive rate by approximately 69%, with a sensitivity decrease of only 4.4 percentage points.

The ablation results in Table 10 show that the final detector performance is not the result of a single threshold. Instead, each component has a distinct role: the score criteria identify strong and amplitude-balanced reflector pairs, the persistence criterion rejects transient responses and the geometry criteria restrict the accepted tracks to physically plausible non-coplanar stepwise configurations.

TABLE 10

ABLATION STUDY OF THE PROPOSED TEMPORAL PAIR-TRACKING DETECTOR

Detector	Sensitivity	False positive rate	Specificity
<i>Score only</i>	93.5%	20.8%	79.2%
<i>Persistence only</i>	93.5%	30.4%	69.6%
<i>Score + persistence</i>	91.3%	9.81%	90.2%
<i>Score + geometry</i>	91.3%	15.1%	84.9%
<i>Final score + persistence + geometry</i>	89.1%	6.5%	93.5%

F. Robustness to scan-region dependence and parameter variation

The robustness of the final detector was evaluated with respect to two possible sources of sensitivity: dependence on individual scan regions and dependence on the selected operating parameters. This analysis was performed after fixing the final detector parameters and was used only to assess the stability of the reported operating point.

The cross-sample holdout stress test showed that selecting the final track-acceptance parameters using only sample S5 led to the same operating point as the final detector listed in Table 3. On the S5 parameter-selection data, this rule detected 32 of 36 expert-labelled positive sequences and produced 23 positive calls among 338 strict-background sequences, corresponding to 88.9% sensitivity and 6.8% false-positive rate. When applied without modification to the holdout S4 data, the same rule detected 9 of 10 expert-labelled positive sequences and produced 4 positive calls among 80 strict-background sequences, corresponding to 90.0% sensitivity and 5.0% false-positive rate. Although the S4 holdout set is small, this result shows that the selected operating point transferred from the larger S5 reference set to the independent S4 shell section without degradation in the observed point estimates.

Scan-region dependence was first assessed using a leave-one-cluster-out analysis. Clusters were defined by sample and circumferential scan index. The detector performance was recalculated after removing one cluster at a time, while keeping the detector parameters unchanged. The resulting performance ranges are summarized in Table 11. For the real HIC-positive group, the full-dataset sensitivity was 89.1% and the leave-one-cluster-out sensitivity varied between 88.4% and 91.1%. The maximum absolute change relative to the full-dataset value was 1.98 percentage points. For the strict-background group, the full-dataset false-positive rate was 6.5% and the leave-one-cluster-out false-positive rate varied between 6.1% and 6.6%. The maximum absolute change was 0.39 percentage points. The results in Table 11 indicate that the reported sequence-level performance was not dominated by a single scan region. Removing individual sample/scan clusters produced only small changes in both sensitivity and false-positive rate, indicating that the reported performance was not driven by any single scan region within the available reference data.

TABLE 11

LEAVE-ONE-CLUSTER-OUT ROBUSTNESS OF THE FINAL DETECTOR

Evaluation group	Full dataset rate	Leave-one-cluster-out minimum	Leave-one-cluster-out maximum	Maximum absolute change
<i>Real HIC-positive sequences, sensitivity</i>	89.1%	88.4%	91.1%	1.98%
<i>Strict-background sequences, false-positive rate</i>	6.5%	6.1%	6.6%	0.39%

A local parameter-variation study was then used to examine whether the selected operating point was isolated or part of a broader stable region. The varied parameters included the minimum track persistence, the maximum and median pair-score thresholds, the maximum allowed median vertical separation and the final median lateral-offset limits. For each parameter combination, the detector was evaluated on the same 46 real HIC-positive sequences and 418 strict-background sequences. The full local sweep contained 1728 tested combinations of the final track-level acceptance parameters. These points were not treated as independent validation experiments, but as a robustness check showing how the fixed detector would behave under local threshold variation. The resulting sensitivity and false-positive-rate trade-off is shown in Figure 10.

The parameter-variation results in Figure 10 show that the final operating point was not an isolated optimum. Several neighbouring parameter combinations maintained sensitivity above 80% while keeping the false-positive rate below 10%. As expected, many parameter combinations fell outside the predefined acceptable region because the sweep deliberately included both stricter and more permissive settings. Nevertheless, 236 combinations satisfied the robustness criterion of sensitivity $\geq 80\%$ and false-positive rate $\leq 10\%$, indicating that the selected operating point was located within the predefined acceptable region and on the Pareto-efficient sensitivity – false-positive-rate frontier. These robustness

analyses should be interpreted within the investigated inspection domain rather than as evidence of universal transferability. The evaluated domain consisted of in-service carbon-steel heat-exchanger shell sections with 46 mm wall thickness, a 7.5 MHz 64-element linear phased-array probe, a Rexolite wedge, the selected sectorial PWI acquisition and TFM reconstruction settings and the VCF×CCF fused image representation. Although the temporal pair-tracking framework is physically interpretable and can be recalibrated, direct transfer of the numerical thresholds to other steel grades, wall thicknesses, probe and wedge configurations, acquisition settings, or HIC morphologies should not be assumed without additional calibration and validation.

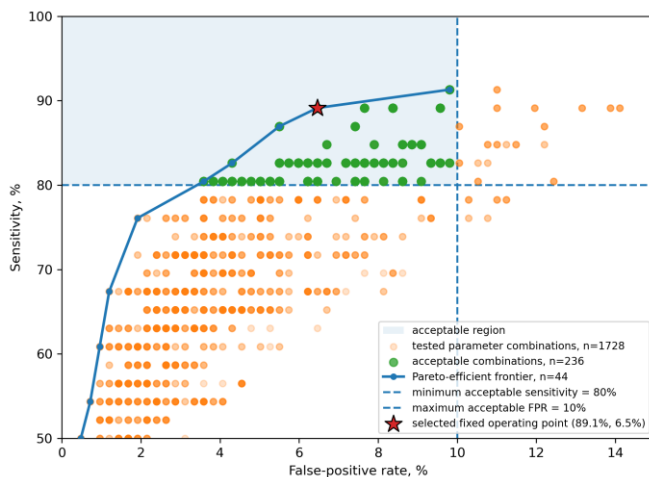


Fig. 10. Local parameter-stability sweep of the final temporal pair-tracking detector. Each point represents one tested combination of the final track-acceptance parameters applied to the same 46 expert-labelled HIC-positive and 418 strict-background sequences. The Pareto-efficient frontier shows operating points for which no other tested parameter combination achieved both lower false-positive rate and higher sensitivity.

IV. CONCLUSION

This study demonstrated a coherence-factor-fusion and physics-constrained temporal pair-tracking approach for detecting stepwise-HIC-like ultrasonic signatures in PWI/TFM images of naturally damaged heat-exchanger shell sections. The VCF×CCF fusion improved the visibility of HIC indications and reduced background artefacts, providing a more suitable image input for automated detection. In particular, the fusion results showed an improvement in contrast-based image quality metrics by 10.9 dB and a reduction of the array performance index from 7.77 to 0.47.

The physical basis of the detection approach was established from metallographic observation of stepwise HIC in sample S5 and from controlled non-coplanar reflector geometries in an artificial calibration block. The calibration-compatible indications, CAL2 and CAL4, were detected by the final fixed detector with an absolute vertical-separation error of 0.1 mm, while the large-offset and boundary cases were rejected under the selected real-HIC operating point. These results confirmed that the detector responds to compact non-coplanar reflector pairs within the targeted sub-millimetre range.

For the detected real HIC-positive sequences, the accepted tracks also provided apparent ultrasonic geometrical

descriptors. The median vertical reflector separation was 0.40 mm, the median lateral offset was 1.00 mm and the median apparent inclination angle was 18.4°. These quantities should be interpreted as image-derived reflector-pair descriptors rather than direct metallographic crack dimensions, but they demonstrate that the proposed detector provides characterization information in addition to a binary sequence-level decision.

When applied to real HIC data, the final detector identified 41 of 46 expert-labelled HIC-positive sequences, corresponding to a sensitivity of 89.1%. The Wilson 95% confidence interval for sensitivity was 77.0–95.3%, while the cluster-bootstrap 95% interval was 79.5–97.7%. In strict-background regions, 27 of 418 sequences were classified as positive, corresponding to a false-positive rate of 6.5% and a specificity of 93.5%. The Wilson 95% confidence interval for the false-positive rate was 4.5–9.2%, while the corresponding cluster-bootstrap interval was 4.2–8.8%. These intervals indicate that the detector maintained a low background response while accounting for both finite sequence count and scan-region grouping. The sensitivity interval remains wider than the false-positive-rate interval because positive stepwise-HIC-like reference sequences were much less numerous than strict-background sequences. Therefore, larger independent positive reference sets would be required for narrow confidence bounds or formal probability-of-detection qualification.

The decision-margin analysis showed clear separation between most HIC-positive and strict-background sequences, while the ablation study confirmed that the low false-positive rate was achieved by combining reflector-pair score, temporal persistence and final geometric constraints. Compared with the score-only rule, the final detector reduced the false-positive rate from 20.8% to 6.5%, with a sensitivity decrease of 4.4 percentage points. This confirms that the final detector performance was not governed by a single amplitude or score threshold, but by the combined physical constraints imposed on the temporal reflector-pair tracks.

Robustness analyses indicated that the reported performance was not dominated by a single scan region or a narrowly tuned parameter set. Leave-one-cluster-out evaluation produced only small performance changes, with a maximum sensitivity change of 1.98 percentage points and a maximum false-positive-rate change of 0.39 percentage points. The local parameter-stability sweep further showed that the selected operating point was located within a broader acceptable region of sensitivity and false-positive-rate performance.

Several limitations remain. Metallographic confirmation was available only for a representative region of Sample S5. The remaining positive sequences are expert NDT reference labels supported by representative metallography and calibration-block evidence. Consequently, the reported sensitivity measures agreement with this reference standard rather than complete diagnostic accuracy against destructive ground truth. Additional targeted metallography is also technically challenging because stepwise HIC is sparse, small and three-dimensional. Finite saw-cut kerf and subsequent material removal during grinding and polishing can cause a nominally targeted section to miss the connecting ligament or intersect only one blister-like segment. The strict-background regions

were likewise not destructively verified as crack-free and the reported false-positive rate should be interpreted as the detector response rate in NDT-defined background regions.

The detector was evaluated here as a sequence-level decision tool, using the strongest accepted temporal track to define the reported decision margin and representative geometry. Spatial redundancy analysis showed that multiple accepted tracks were uncommon: 38 of 41 detected positive sequences contained one accepted track, three contained two accepted tracks and none contained three or more. Therefore, the reported sequence-level performance was not inflated by repeated counting of multiple tracks within the same local window. Nevertheless, the current implementation does not yet provide a full indication-level de-duplication or multi-defect separation stage. Consequently, two closely spaced physical HIC clusters could be represented as one sequence-level indication if their ultrasonic responses satisfy the same local pairing and temporal-linking constraints. Future work should extend the output stage using spatial non-maximum suppression or clustering of accepted track centres to merge redundant tracks and to report multiple spatially separated stepwise-HIC-like indications individually.

The detector was evaluated primarily with a fixed operating point tied to the selected VCF×CCF image representation. The fusion-comparison analysis showed that changes in image morphology can substantially affect sensitivity and false-positive rate. However, even when the final track-acceptance thresholds were optimized independently for each fusion method, VCF×CCF retained the highest maximum Youden index and the most favourable sensitivity–false-positive-rate balance. The reported operating point should therefore be considered specific to the present inspection domain: in-service carbon-steel heat-exchanger shell sections with 46 mm wall thickness, the 7.5 MHz phased-array probe and Rexolite wedge configuration, the selected sectorial PWI acquisition and TFM reconstruction settings and the VCF×CCF fused image representation. Although the temporal pair-tracking framework is physically interpretable and can be recalibrated, direct transfer of the numerical thresholds to other steel grades, wall thicknesses, probe configurations, wedge geometries, acquisition settings or HIC morphologies should not be assumed without additional calibration and validation.

Nevertheless, the overall results show that combining coherence-factor fusion with an interpretable temporal pair-tracking detector can support the identification of stepwise-HIC-like indications in ultrasonic inspection data, while keeping each decision traceable to explicit physical criteria rather than to a black-box model. The proposed detector is not intended to replace expert evaluation, but to provide a physically constrained sequence-level decision and decision margin that can help prioritize indications for further review.

ACKNOWLEDGMENT

This work was supported by the Research Foundation of the Research Council of Lithuania under the project CURIOS: „Machine learning assisted detection and grading of hydrogen-induced cracking using super-resolution ultrasonic phased array imaging“, grant number: S-MIP-24-45.

REFERENCES

- [1] Shang J, Lu Y, Zheng J, Sun C, Hua Z, Yu W, et al. Research status-in-situ and key challenges in pipeline transportation of hydrogen-natural gas mixtures. *Huagong Jinzhan/Chemical Industry and Engineering Progress* 2021;40(10):5499-505.
- [2] Martin P, Ocko IB, Esquivel-Elizondo S, Kupers R, Cebon D, Baxter T, et al. A review of challenges with using the natural gas system for hydrogen. *Energy Sci.Eng.* 2024;12(10):3995-4009.
- [3] Zhang F, Zhao P, Niu M, Maddy J. The survey of key technologies in hydrogen energy storage. *Int J Hydrogen Energy* 2016;41(33):14535-52.
- [4] Islam A, Alam T, Sheibley N, Edmonson K, Burns D, Hernandez M. Hydrogen blending in natural gas pipelines: A comprehensive review of material compatibility and safety considerations. *Int J Hydrogen Energy* 2024;93:1429-61.
- [5] Chang W, Wang W, Guo Y, Wang L, Li Z. Diffusion characterization of hydrogen-blended natural gas leakage for buried pipeline based on simulation. *Int J Hydrogen Energy* 2025;99:394-408.
- [6] Qin G, Li R, Wang B, Wang Y, Cheng YF. Enhancing hydrogen pipeline safety: Perspectives on resilience against hydrogen-induced degradation. *Journal of Pipeline Science and Engineering* 2025:100314.
- [7] Ohaeri E, Eduok U, Szpunar J. Hydrogen related degradation in pipeline steel: A review. *Int J Hydrogen Energy* 2018;43(31):14584-617.
- [8] Mohtadi Bonab MA, Szpunar J, Basu R, Eskandari M. The mechanism of failure by hydrogen induced cracking in an acidic environment for API 5L X70 pipeline steel. *Int J Hydrogen Energy* 2015;40.
- [9] Tetelman AS, Robertson WD. Institute of Metals Division - The Mechanism of Hydrogen Embrittlement Observed in Iron-Silicon Single Crystals. 1962;.
- [10] Xu T, Xiong F, Han H, Rong H, Tang J, Jiang H, et al. Research on hydrogen-induced crack propagation behavior in the girth weld zone of X80 hydrogen-enriched pipelines based on the phase field method. *Eng. Failure Anal.* 2025;174:109528.
- [11] Martin ML, Sofronis P. Hydrogen-induced cracking and blistering in steels: A review. *Journal of Natural Gas Science and Engineering* 2022;101:104547.
- [12] Elboujdaini M. Hydrogen-Induced Cracking and Sulfide Stress Cracking. *Uhlig's Corrosion Handbook*; 2011, p. 183-194.
- [13] Lozev M, Neau GA, Yu L, Eason T, Orwig SE, Collins RC, et al. Assessment of High-Temperature Hydrogen Attack Using Advanced Ultrasonic Array Techniques. *Mater.Eval.* 2020:1223-38.
- [14] Spies M, Rieder H, Dillhöfer A, Schmitz V, Müller W. Synthetic aperture focusing and time-of-flight diffraction ultrasonic imaging - Past and present. *Journal of Nondestructive Evaluation*, 2012;31(4).
- [15] Nugent M, Silfies T, Dobis J, Armit T. A Review of High Temperature Hydrogen Attack (HTHA) Modeling, Prediction, and Non-Intrusive Inspection in Refinery Applications. *CORROSION* 2017, New Orleans, Louisiana, USA, March 2017.
- [16] Birring AS, Riethmuller M, Kawano K. Ultrasonic techniques for detection of high temperature hydrogen attack. *Mater.Eval.* 2005;63:110-5.
- [17] Trimbom N. Detecting and Quantifying High Temperature Hydrogen Attack (HTHA) . 2016.
- [18] Nageswaran C. Maintaining the integrity of process plant susceptible to high temperature hydrogen attack. Part 1: analysis of non-destructive testing techniques . TWI Ltd 2018;.
- [19] Le Nevé Ch, Loya S, Le Jeune L, Mahaut S, Demonte S, Chauveau D, Renaud R, Tessier M, Nourrit N, Le Guellaut A. High Temperature Hydrogen Attack: New NDE Advanced Capabilities — Development and Feedback. *ASME 2019 Pressure Vessels & Piping Conference* July 14–19, 2019, San Antonio, Texas, USA.
- [20] Kažys RJ, Mažeika L, Samaitis V, Šlitis R, Merck P, Viliūnas Ž. Comparative Analysis of Ultrasonic NDT Techniques for the Detection and Characterisation of Hydrogen-Induced Cracking. *Materials* 2022;15(13).
- [21] Dupont-Marillia F, Krynicki JW, Bélanger P. Early detection of high temperature hydrogen attack using the ultrasonic full matrix capture and advanced post-processing methods. *NDT E Int.* 2022;130:102676.
- [22] Le Jeune L, Robert S, Villaverde EL, Prada C. Multimodal Plane Wave Imaging for Non-destructive Testing. *Physics Procedia* 2015;70:570-3.
- [23] Le Jeune L, Robert S, Lopez Villaverde E, Prada C. Plane Wave Imaging for ultrasonic non-destructive testing: Generalization to multimodal imaging. *Ultrasonics* 2016;64:128-38.

- [24] Holmes C, Drinkwater BW, Wilcox PD. Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. *NDT E Int.* 2005;38(8):701-11.
- [25] Prado VT, Granja SCG, Higuti RT, Kitano C, Martínez-Graullera O, Segura LE. Defect detection in anisotropic plates based on the instantaneous phase of signals. *IEEE Trans.Ultrason.Ferroelectr.Freq.Control* 2015;62(10):1888-94.
- [26] Gauthier B, Painchaud-April G, Le Duff A, Bélanger P. Ultrasonic multi-view data merging using the vector coherence factor. *NDT E Int.* 2023;136:102805.
- [27] Lin J, editor. Ultrasonic array imaging in nondestructive evaluation: total focusing method with using circular coherence factor. ; 2013/01/1; ; 2013.
- [28] Camacho J, Fritsch C, Fernandez-Cruza J, Parrilla M. Phase Coherence Imaging: Principles, applications and current developments. *Proc.Mtgs.Acoust.* 2020;38(1):055012.
- [29] Cruza JF, Camacho J, Fritsch C. Plane-wave phase-coherence imaging for NDE. *NDT E Int.* 2017;87:31-7.
- [30] Camacho J, Atehortua D, Cruza JF, Brizuela J, Ealo J. Ultrasonic crack evaluation by phase coherence processing and TFM and its application to online monitoring in fatigue tests. *NDT E Int.* 2018;93:164-74.
- [31] Lesage JC, Marvasti M, Farla O. Vector coherence imaging for enhancement of small omni-directional scatterers and suppression of geometric reflections. *NDT E Int.* 2021;123:102502.
- [32] Luo L, Tan Y, Li J, Zhang Y, Gao X. Wedge two-layer medium ultrasonic plane wave compounding imaging based on sign multiply coherence factor combined with delay multiply and sum beamforming. *NDT E Int.* 2022;127:102601.
- [33] Zhou J, Xue J, Bi X, Hu N. Angle weighted sign coherence factor-based ultrasonic fusion imaging for crack detection. *Measurement* 2025;253:117594.
- [34] Fei Q, Cao J, Xu W, Jiang L, Zhang J, Ding H, et al. A Deep Learning-Based Ultrasonic Diffraction Data Analysis Method for Accurate Automatic Crack Sizing. *Applied Sciences* 2024;14(11).
- [35] Niu S, Srivastava V. Simulation trained CNN for accurate embedded crack length, location, and orientation prediction from ultrasound measurements. *Int.J.Solids Structures* 2022;242:111521.
- [36] El-Hawwat S, Shah JK, Wang H, Venkateela G. Detection of internal cracks in polyethylene pipes using ultrasonic imaging and deep learning. *Measurement* 2025;253:117491.
- [37] J. Camacho, M. Parrilla, C. Fritsch. Phase Coherence Imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 2009;56(5):958-74.
- [38] Tian J, Chen Z, Lu S, Qi X, Pu S, Liu M. Phase coherence weighted ultrasound total focusing method towards the improved imaging of CFRP defects. *Composites Communications* 2023;43:101736.
- [39] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 1979;9(1):62-6.
- [40] ASTM, E2737-10. Standard Practice for Digital Detector Array Performance Evaluation and Long-Term Stability. 2018;.
- [41] Usamentiaga R, Ibarra-Castaneda C, Maldague X. More than Fifty Shades of Grey: Quantitative Characterization of Defects and Interpretation Using SNR and CNR. *J.Nondestr.Eval.* 2018;37(2):25.
- [42] R. K. Rachev, P. D. Wilcox, A. Velichko, K. L. McAughy. Plane Wave Imaging Techniques for Immersion Testing of Components With Nonplanar Surfaces. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 2020;67(7):1303-16.



Vyintas Samaitis (Member, IEEE) received the M.Sc. degree in 2012 and the Ph.D. degree in Electrical and Electronics Engineering in 2016. Since 2026, he has been a Chief Researcher at the Prof. K. Baršauskas Ultrasound Research Institute, Kaunas University of Technology, where he also heads the Laboratory of Ultrasonic Measurements. His research interests include long-range guided-wave inspection, structural health monitoring,

material characterization, signal processing, phased-array imaging and machine learning. He is the author or co-author of more than 40 scientific papers and co-author of two national patents and one European patent. He is a member of IEEE, ESIS, the Lithuanian Scientific Society and the Lithuanian Society for Non-Destructive

Testing and Technical Diagnostics. He is also an expert for the Research Council of Lithuania and the Research Council of Norway.



Renaldas Raišutis (Member, IEEE) received the M.S. degree in Ultrasound Technology in 2001 and the Ph.D. degree in Measurement Engineering in 2005. Since 2014, he has been a Chief Researcher and Head of the Numerical Simulation Laboratory at the Prof. K. Baršauskas Ultrasound Research Institute, Kaunas University of Technology, where he has served as Director since 2024. He is also a Professor in the Department of Electrical Power Systems, Faculty of Electrical and Electronics Engineering, Kaunas University of Technology. His research interests include fundamental and applied ultrasound, non-destructive testing, monitoring and quality control, signal and image processing, material characterization, clinical decision-support solutions, predictive maintenance and diagnostics. He is the author or co-author of more than 100 scientific publications and co-author of three national patents, seven international patents and nine patent applications. He is a member of IEEE, ESIS and EFSUMB. He also serves as an expert and consultant for local and international industry, start-ups and research organizations.



Audrius Stravinskas received the M.Sc. degree in Ultrasound Technology from Kaunas University of Technology, Lithuania, in 2001. He has broad professional experience in software development, electronic device design and programming. Since 2023, he has been a Ph.D. student at Kaunas University of Technology, focusing on ultrasound-based non-destructive evaluation of hydrogen-induced cracking and creep damage. His research interests include ultrasonic signal and image processing,

advanced non-destructive testing techniques and the integration of machine-learning methods for improved defect detection and defect-type characterization.



Aadhik Asokkumar received the M.Sc. degree from Dresden International University, Germany, in 2018 and the Ph.D. degree in Measurement Engineering from Kaunas University of Technology, Lithuania, in 2025. He is currently a Researcher at the Prof. K. Baršauskas Ultrasound Research Institute, Kaunas University of Technology. From 2019, he served as a Junior Researcher at the same institute, contributing to several international and national projects, including the Horizon Europe

Euratom projects FIND and iWeld, the Research Council of Lithuania project CURIOUS and the Marie Skłodowska-Curie networks GW4SHM and NDTonAIR. His research interests include ultrasonic testing, guided-wave-based structural health monitoring, signal processing, damage imaging and finite-element simulation for ultrasonic testing.