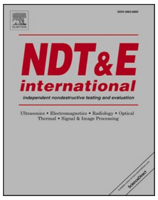
NDT&E International 144 (2024) 103094

Contents lists available at ScienceDirect

NDT and E International

journal homepage: www.elsevier.com/locate/ndteint



A study of machine learning object detection performance for phased array ultrasonic testing of carbon fibre reinforced plastics

Vedran Tunukovic [a,b,*], Shaun McKnight [a], Ehsan Mohseni [a], S. Gareth Pierce [a], Richard Pyle [a], Euan Duernberger [a], Charalampos Loukas [a], Randika K.W. Vithanage [a], David Lines [a], Gordon Dobie [a], Charles N. MacLeod [a], Sandy Cochran [b], Tom O'Hare [c]

[a] *Sensor Enabled Automation, Robotics, and Control Hub (SEARCH), Centre for Ultrasonic Engineering (CUE), Electronic and Electrical Engineering Department, University of Strathclyde, Glasgow, UK*
[b] *Future Ultrasonic Engineering, FUSE CDT, Glasgow, UK*
[c] *Spirit AeroSystems, Belfast, UK*

ARTICLE INFO

Keywords:
Automated defect detection
Machine learning
Deep learning in ultrasonic testing
Non-destructive evaluation
Carbon fibre reinforced polymers
Aerospace composites

ABSTRACT

The growing adoption of Carbon Fibre Reinforced Plastics (CFRPs) in the aerospace industry has resulted in a significant reliance on Non-Destructive Evaluation (NDE) to ensure the quality and integrity of these materials. The interpretation of large amounts of data acquired from automated robotic ultrasonic scanning by expert operators is often time consuming, tedious, and prone to human error creating a bottleneck in the manufacturing process. However, with ever growing trend of computing power and digitally stored NDE data, intelligent Machine Learning (ML) algorithms have been gaining more traction than before for NDE data analysis. In this study, the performance of ML object detection models, statistical methods for defect detection, and traditional amplitude thresholding approaches for defect detection in CFRPs were compared. A novel augmentation technique was used to enhance synthetically generated datasets used for ML model training. All approaches were tested on real data obtained from an experimental setup mimicking industrial conditions, with ML models showing improvement over amplitude thresholding and statistical thresholding techniques. The advantages and limitations of all methods are reported and discussed.

1. Introduction

Composites are defined as materials comprising two or more distinct natural or synthetic constituents that exhibit a unique set of mechanical characteristics. Their utilisation has been prominent in various fields such as renewable energy, aerospace, construction, sports equipment, biomedical field, automotive, and marine industries.

Carbon Fibre Reinforced Plastics (CFRPs) are constructed of structured carbon fibre sheets bonded together by a polymer matrix and are widely used type of composite material in the aerospace industry. This mechanism of construction offers several benefits, including improved corrosion and fatigue resistance, high specific strength, and lightweight structure [1,2]. The rising trend of CFRPs uptake in the aerospace industry is particularly driven by the improved fuel efficiencies achieved by lower overall aircraft weight [3]. Currently, composites constitute approximately 50–53% of the structural mass of flagship aircraft from Airbus and Boeing, such as the A350XWB and 787 Dreamliner [4,5]. This considerable use of composites calls for a thorough post-manufacturing inspection of all produced CFRP parts, as they are susceptible to a variety of defects such as delaminations, inclusions, and porosities [6–9]. CFRPs are used to construct critical components such as fuselage, wing covers, engine covers, and stabilizers, making these material imperfections a threat to the mechanical and structural integrity of the aircraft.

Non-Destructive Evaluation (NDE) is an umbrella term for various methods that are used to test and evaluate engineering systems and materials without causing damage to the inspected components. Some of the processes used include radiographic testing, thermographic testing, visual inspection, and ultrasonic testing.

Ultrasonic Testing (UT) is the most used bulk inspection method in

[*] Corresponding author. Sensor Enabled Automation, Robotics, and Control Hub (SEARCH), Centre for Ultrasonic Engineering (CUE), Electronic and Electrical Engineering Department, University of Strathclyde, Glasgow, UK.
E-mail address: vedran.tunukovic@strath.ac.uk (V. Tunukovic).

https://doi.org/10.1016/j.ndteint.2024.103094
Received 5 July 2023; Received in revised form 27 November 2023; Accepted 9 March 2024
Available online 15 March 2024

aerospace due to its flexibility, ease of use, and safety [9,10]. During the testing process, an ultrasonic transducer generates acoustic waves that propagate through the material, interact with the test structure, and are reflected/diffracted by boundaries and discontinuities. The returning waves are modified by factors such as attenuation, scattering, absorption, and inter-layer/inter-material boundaries, as well as potential defects. The returning waves are received by the same transducers, converted to voltage due to the piezoelectric effect, and recorded as time-series data, providing information about the internal structure of the material. Phased Array Ultrasonic Testing (PAUT) is a variant of UT that uses multiple ultrasonic transducers/elements to enable more complex material scanning by introducing calculated time delays in the transmission/reception of individual elements. PAUT allows for the use of techniques such as beamforming, beam steering, linear scanning, dynamic focusing, and full matrix capture [11–13]. Data collected with PAUT is usually displayed as a B-scan, a 2D intensity map representation of time-series data acquired by a group of elements, or as an amplitude C-scan, a cross-sectional representation of gated time-series data that highlights the highest amplitude responses.

The introduction of robotic automation in the field of NDE has led to a significant reduction in inspection time compared to a manual approach. In addition, the use of robotic inspection diminishes the need for human labour, resulting in a precise and repeatable evaluation process [14]. However, the automation of NDE procedures has also increased the volume of data that must be analysed. The interpretation of collected information is done manually by an NDE operator, which presents a continuous challenge as it is time-consuming and exhibits the potential for human error [15,16]. While robotic automation accelerates the NDE scanning processes, the speed of data interpretation remains constrained by a manual procedure, making the development of a robust automated or human augmented data interpretation system a desirable addition to the NDE process.

Machine learning (ML) is a subcategory of Artificial Intelligence (AI), an umbrella term that encompasses the science of developing computers, robots, and other devices capable of performing various tasks with the equivalent of human proficiency. The idea behind ML is to create and develop algorithms and methods that improve over time based on data. Deep Learning (DL) is a subcategory of ML that is focused on developing more complex algorithms, taking inspiration from the human brain.

In the past decade, the number of academic publications on the application of ML technology in NDE has increased significantly. When it comes to UT, ML is extensively used on data collected from bulk metallic materials and welds [17–29]. ML models were used in Ref. [17] to determine fatigue life and tensile strength of welds from ultrasonic A-scans. The main achievement was the correlation of mechanical properties to raw UT data which in turn enables predictions of mechanical properties that require destructive methods without damaging the components. In a series of papers [22–24], researchers explored ML models for classification of weld flaws. The progression of the work was characterised by the overall increase in the performance by adoption of deeper neural networks, various augmentation techniques, and denoising autoencoders to improve quality of the inputs. Similarly, the works presented in Refs. [16,19] focused on augmentation of data to improve UT scans of austenite welds. A proprietary software for augmentation of defective signals was used to drive the training of ML networks that ultimately outperformed NDE inspectors. Authors have compared different feature extraction techniques in Ref. [21] to generate inputs for ML classifiers. They have explored various time to frequency domain transforms in order to successfully classify UT A-scan data. Density-based spatial clustering of applications with noise algorithm were used in Ref. [30] to cluster healthy and unhealthy signals. Authors demonstrated promising results with features extracted directly from the raw A-scans. In Ref. [31], Convolutional Neural Network (CNN) that can determine crack dimensions, location, and orientation in load-bearing structures was developed. The training was based on A-scan data created with Finite Element Analysis (FEA) software and authors demonstrated good generalisation to experimental data.

In contrast, a relatively smaller body of work has focused on composites [32–36]. Authors in Ref. [32] compared the classification performance of multiple models and feature extractors on A-scan data captured from CFRPs with manually embedded defects. They have concluded that CNNs were the best feature extractors for this application. Work conducted in Ref. [33] focused on the development of a fully convolutional neural network that classified A-scans collected from a 3D braided composite. In this approach, researchers have used a single A-scan as an input to determine the presence of debonding in their sample. Convolutional autoencoders were used in Ref. [34] to determine fatigue damage with ultrasonic-guided wave imaging while highlighting the issue of gathering data to drive model training. Various signal decomposition techniques were compared in Ref. [35] to improve the feature extraction process for ML training as defects in composites are oftentimes masked by larger features such as front and back wall reflections. Lastly, in Ref. [36] the authors employed ML and guided waves to assess damage in composite structures, achieving promising results despite the introduction of various influencing factors, such as different temperatures.

In recent years, there has been an abundance of development of new object detection models with the examples being R–CNN (Region-based Convolutional Neural Networks), Fast R–CNN, Faster R–CNN, EfficientDet, and You Only Look Once (YOLO) [37–43]. These models use a complex architecture to extract regions of interest of an input image, outputting both the area of interest and class of the object in the form of a vector. Despite the rise in the number of publications, object detection models have seen limited implementation with UT data. Performances of EfficientDet, RetinaNet, and YOLOv5 models on volume-corrected B-scans of steel samples were compared in Ref. [25]. Authors have reported promising results with architectural changes made to address the issue of extreme aspect ratios observed in UT B-scans. Similarly, object detection on ultrasonic B-scans was evaluated in Ref. [29], demonstrating the use of YOLO and Single Shot Detector models and highlighting the differences in performance in inference speed between the tested models. Lastly, researchers in Ref. [27] combined EfficientDet and several methods that enabled processing of additional B-scans in the sequence, improving on the baseline results.

In industrial applications, defect localization and sizing are usually performed manually through visual inspection of the C-scan, while applying different thresholds to the image. The most used method is a 6 dB drop where a threshold value is imposed to the signal to separate healthy and potentially defective regions. Researchers have used 6 dB drop to separate damaged and undamaged areas in a C-scan image to assess the extent and size of impact damage [44]. The authors compared how sizing results vary with different methods and proposed a new algorithm that improves sizing and shape of the damage. Limitations of 6 dB method was recognized in Ref. [45], especially when sizing defects that are smaller than the width of the ultrasonic beam. As an improvement the authors developed an ML approach that can automatically acquire different thresholding values, hence reducing the errors in quantification of defects. A semi-automated detection algorithm was proposed in Ref. [46]. This approach works on time-of-flight C-scans, where user defines areas of interest and threshold values which are in turn used for automated analysis.

Automatic defect localization in CFRPs has been scarcely explored in the past; authors of [47] developed a time-dependent thresholding that improved detection of micro flaws in ultrasonic C-scans of stainless-steel samples. Statistical analysis of backscattering noise to determine defect locations was used in Ref. [48], but the scope of their work was limited. Several works have used Otsu thresholding to segment ultrasonic images into clusters of areas with similar acoustic properties [49–51]. The most recent work was presented in Ref. [52], where an ML object detection model successfully localized damage on time-of-flight C-scans of aircraft wings. The authors demonstrated accuracy of 94.5% for the best performing model when training and testing on experimentally collected data.

Provided the limited past research investigations, and the broad gap in the knowledge for automated defect detection, this work for the first time presents a comparison between the capability of various defect detection methodologies. Firstly, an amplitude thresholding method, frequently used within the industry, was trialled as a baseline for comparison. Afterwards, an improvement was shown for the thresholding with the implementation of statistical amplitude thresholding method, inspired by previous work in fusion of ultrasonic data [53]. Lastly, the reliability of ML algorithms based on widely used object detection models such as YOLO, Faster R–CNN, and RetinaNet was investigated with slight modifications to the architectures and the key strengths and shortcomings were highlighted. The training datasets for ML models were created with semi-analytical software and augmented with methods proposed by Ref. [54]. Consequently, dependence on experimental data is diminished as large volumes of real defect responses in CFRPs are not readily available. These combined methods allowed for development a robust ML process that does not require inclusion of experimental data during the training. All the methods were tested on a series of CFRP samples with varying characteristics and embedded defects that were scanned with state-of-the-art automated robotic PAUT roller probe scanning setup that mimics the setups currently used in industry.

The rest of the paper is organised as follows: Section 2. covers materials and methods that were used, Section 3. provides results and a discussion of all evaluated methods. Lastly, Section 4. provides the conclusions and the prospects for future work.

2. Materials and methods

2.1. Acquisition of experimental data

Five CFRP samples manufactured to a BAPS 260 standard were supplied by Spirit AeroSystems. These samples were produced using the resin infusion method, woven fabric sheets and Cycom 890 polymer. To capture acoustic responses similar to those produced by defects, Flat Bottom Holes (FBHs) were fabricated in two samples and rectangular-shaped Teflon and bagging film inserts were introduced to the other two samples. One of the samples was kept as a control without any defects. FBHs and Teflon inserts are often used to mimic the acoustic responses of the delamination that can occur during manufacturing processes [55]. According to the internal Spirit AeroSystems' current guidelines for NDE inspection (internal document, not publicly accessible), critical defect sizes are described according to their type and location on the aircraft. For delaminations, the largest allowable flaw area that would not be categorised as defect, ranges from 60 to 500 mm^2, depending on the location on the aircraft. This range represents quite large defects that normally can be easily spotted on C-scan images. However, to challenge and understand the limits of the defect detection algorithms and PAUT inspection setup, FBHs with diameters ranging from 3.0 to 9.0 mm, and square Teflon and bagging film inserts with dimensions of 4.0–12.0 mm were embedded into CFRP samples. This was done to scrutinise the performance of defect detection algorithms and test flaws with areas between 7.0 and 144.0 mm^2.

For sample A with dimensions of 254.0 mm × 254.0 mm x 8.6 mm, the diameter of drilled FBHs were 3.0, 6.0, and 9.0 mm respectively with a ± 0.2 mm tolerance at depths of 1.5, 3.0, 4.5, 6.0, and 7.5 mm with ± 0.3 mm tolerance measured from the front face of the sample. FBHs were spaced 35 mm and 30 mm apart on X and Y axis, respectively. In addition to these 15 manufactured FBHs, an in-depth analysis of B and C-scans confirmed the presence of two additional smaller delaminations. Sample B with dimensions of 254.0 mm × 254.0 mm x 8.6 mm was made similarly, with the addition of 4.0 and 7.0 mm FBHs, resulting in 25 defects in total. 1.0 mm diameter FBHs were also trialled in the study however, the current measurement setup was unable to capture them. Sample C of dimensions 780 mm × 200 mm with 5 different thicknesses ranging from 7.5 mm to 16.0 mm in steps of 2.1 mm was also used in the study. Each thickness step had 3 embedded Teflon and bagging film inserts of sizes 9.0 and 6.0 mm, with two of them being positioned immediately subsurface, two in the middle of the sample, and two close to the back wall of the sample. Sample D was smaller, with dimensions of 300 mm × 90 mm with a total of 14 embedded rectangular Teflon tapes of sizes 12.0, 6.0, and 4.0 mm. The final 254.0 mm × 254.0 mm x 8.6 mm sample E did not have any defects and was used to produce pristine scans and for synthetic dataset augmentation explained in section 2.4. The second sample with drilled FBHs and CAD drawing is presented in Fig. 1 and the summary of all used samples is presented in Table 1.

The experimental system setup was established based on the previous work presented in Ref. [56] to enable automated deployment of the ultrasonic probe and registration of robotically encoded inspection data. This was achieved using a Kuka KR90 R3100 extra HA industrial manipulator [57] with 6 degrees of freedom. The manipulator's maximum reach of 3095 mm and a maximum payload of 90 kg, combined with a pose repeatability of ± 0.04 mm, ensured that measurements were precise and consistent.

Path planning was performed within a central LabVIEW Virtual Instrument (VI) control program on a desktop PC connected through Ethernet to the robot and the phased array controller. An Inspection Solutions RollerFORM-5L64 [58], encasing a 5 MHz array of 64 elements with 0.8 mm pitch and 6.4 mm elevation, by Olympus NDT was used for the robotic experiments as the probe geometry with its rolling capability is tailored for easy integration with robotic manipulators. To improve coupling and wave propagation, the tyre is made from low-attenuation material with a similar acoustic impedance to water. The interior of the tyre was filled and pressurized with glycol to prevent the formation of air bubbles. The roller probe was mounted on the KUKA KR90 3100 extra HA, which enabled programmatic movement at a constant speed of 10 mm/s. During the acquisition, water was used as a couplant between the exterior of the tyre and the inspected CFRP sample.

Even though the probe is used on the surface with sprayed water coupling between its tyre and the component's surface, achieving stable and constant contact force is crucial to sustained image quality during the mechanical scan. Therefore, real-time corrections and control were accomplished for the PAUT probe movement normal to the component's surface to maintain a constant coupling force throughout the surface raster scan. The real-time vertical position control was enabled through the KUKA RobotSensorInterface software package and an adaptive force-torque motion control program created within the central LabVIEW VI and based on the real-time measurements of a Schunk GmbH & Co. FTN-GAMMA-IP65 SI-130-10 Force Torque (FT) sensor mounted between the probe and the robot's end effector [59]. FT sensor enabled 3-dimensional measurements of forces and torques, within a range of 400 N in vertical (z) and 130 N in horizontal (x,y) directions. FT also served as a fail-safe measure programmed to stop the movement of the industrial manipulator if the contact force exceeds a preset value; which was set to 150 N to protect the PAUT roller-form. The industrial manipulator, FT sensor, and ultrasound roller probe assembly are illustrated in Fig. 2.

A MicroPulse 6 [60] controller, by Peak NDT Ltd., with 128 transmission and reception (T/R) channels with a maximum pulse voltage of 200 V was used to drive the Olympus phased array. The array was excited in linear electronic scanning mode with a sub-aperture of 4 elements, an excitation voltage of 80 V, a reception gain of 22.5 dBs, and a pulse width of 100 ns. A Digital 6 MHz lowpass filter was used to filter out unwanted higher frequency signals that might induce resonance of near-surface carbon fibre layers [61]. For data acquisition, scanning speed of 10 mm/s was used with the pulse repetition rate of 760 Hz. The digitiser of MicroPulse 6 was set up to capture data at a sample rate of 100 MHz in 32-bit precision. The T/R instructions for the phased array controller were written in a MicroPulse command file format containing

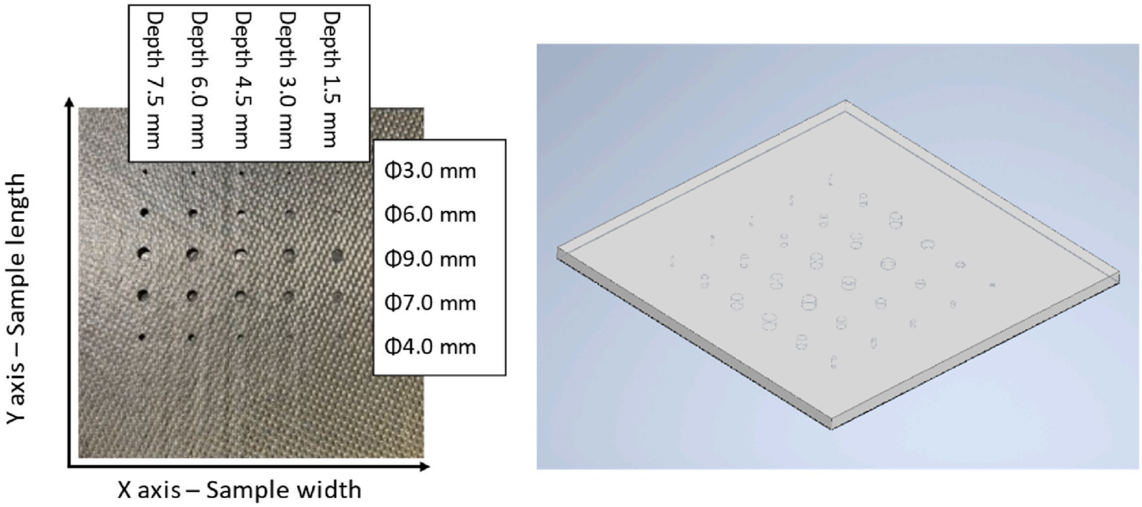


Fig. 1. Sample B (left) and 3D CAD model of the same sample (right).

Table 1
Summary of used samples.

Sample	Dimensions [mm]	Thickness [mm]	Defects
A	254.0 × 254.0	8.6	15 FBHs and 2 delaminations
B	254.0 × 254.0	8.6	25 FBHs
C	780.0 × 200.0	7.5–16.0	12 Teflon and 12 bagging film inserts
D	300.0 × 90.0	8.0–21.0	14 Teflon inserts
E	254.0 × 254.0	8.6	N/A - Pristine sample

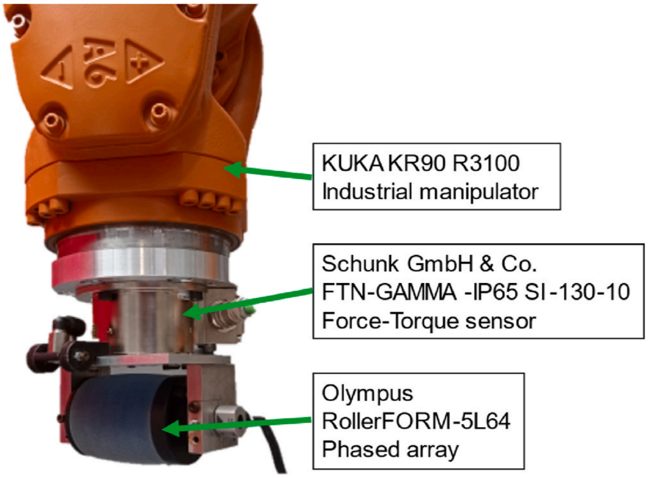


Fig. 2. Assembly of the industrial manipulator, force-torque sensor, and phased array roller probe used in the study.

the instructions on the operating sequence and properties of individual array elements. To account for the depth-wise loss of amplitude of transmitted acoustic wave, time compensated gain was set up in the form of a linear ramp function, starting at 0 dB of additional gain at 1.5 mm depth and ending at 23.75 dB at a depth of 15 mm. The ramp was determined experimentally in a way so that the front and back-wall acoustic responses match in amplitude. The system architecture is shown in Fig. 3, where green and blue blocks represent software and hardware components respectively.

The robotic path was adjusted with graphic user interface designed in LabView software. With 4 element sub-aperture, the active aperture equalled 48.8 mm, so multiple robotic passes with an offset of 48.8 mm were performed to create a rasterized scan of the sample. The base position of the scan start was kept constant for all the scanned samples. Illustrations of rasterized path planning and example of a C-scan image are presented in Fig. 4.

2.2. *Generation of simulated data*

For the generation of simulated data, a semi-analytical NDE UT software CIVA by EXTENDE S.A. was used [62]. This approach is more efficient and less computationally demanding than using the Finite Element Analysis (FEA) software, as it uses ray tracing theory. However, this assumption does not yield results that represent the properties of real-world composite materials accurately as it undermines the noise levels caused by the inter-laminar scattering and diffraction. FEA software, on the other hand, can simulate and calculate complex interactions between the wave and individual layers of the composite with greater accuracy, resulting in a more representative simulation. However, this often requires painstaking definition of individual layer's properties and dimensions. To gain an understanding of simulation time differences between CIVA and FEA, a similar scenario of a probe on a defective sample was modelled in CIVA and an FEA wave propagation software POGO [63]. Both simulations were executed on a high-performance PC with Intel® Xeon(R) Gold 6248R Central Processing Unit (CPU), Nvidia RTX 3090Ti Graphics Processing Unit (GPU), and 192 GB Random Access Memory (RAM). The CPU-intensive CIVA simulation was completed in less than 2 min, whereas POGO FEA simulations even with GPU parallelisation took more than 2 h of processing. In this study, it was decided to create 300 simulations of defect responses to include a variety of defect sizes at different depths in the CFRP sample. Given the large number of simulations and the simulation time discrepancy observed between the CIVA and FEA, it was decided to use a semi-analytical modelling approach and attempt to reintroduce the compromised signal features in post-processing stage.

Upon deciding the simulation software, a square composite sample with dimensions of 100 mm × 150 mm x 8 mm was created and a range of FBHs were introduced in the model. A parametric sweep study was used for ease of data collection, where FBHs' diameters ranged from 3.0 to 15.0 mm, each placed at depths of 1.5–7.5 mm in steps of 0.5 mm measured from the inspection surface. Each simulation in the sweep contained only 1 defect in the centre of the sample. The flow chart of the simulation process and an example output data for a defect of 6.0 mm at the depth of 4.5 mm is displayed in Fig. 5 in form of a C-scan.

The composite model was defined with a total of 8 carbon fibre layers in orientations of 0°, 45°, 90°, −45°, 0°, 45°, 90°, and −45° with the

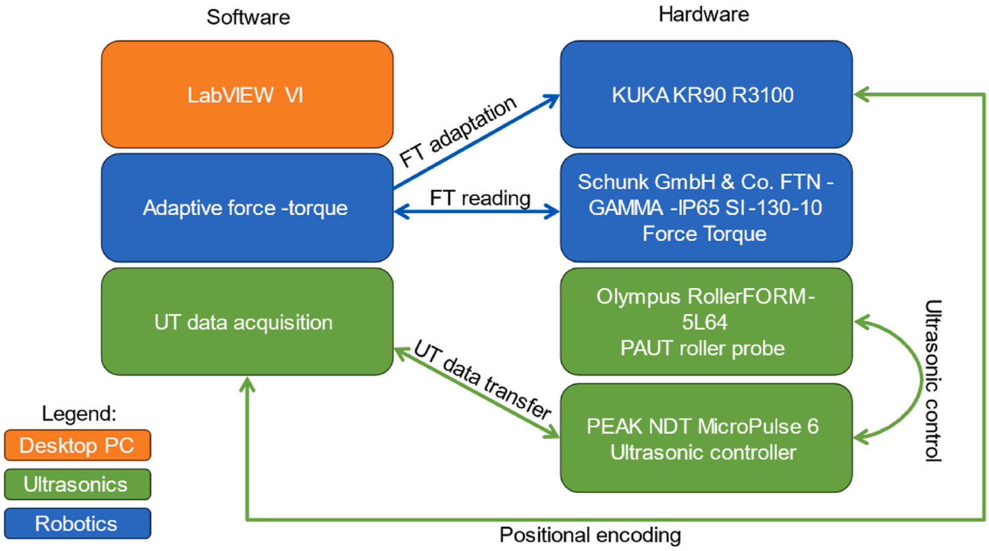


Fig. 3. System design of sensor-enabled robotic scanning with ultrasonic phased array roller probe.

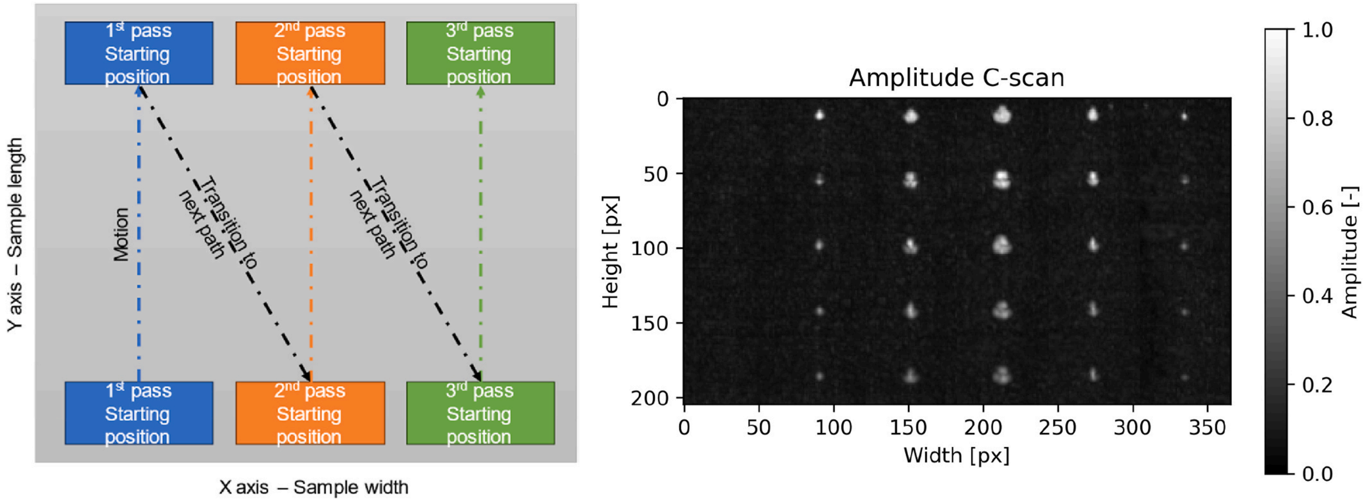


Fig. 4. Robotic path planning for raster scan (left) and resulting C-scan image of sample B (right).

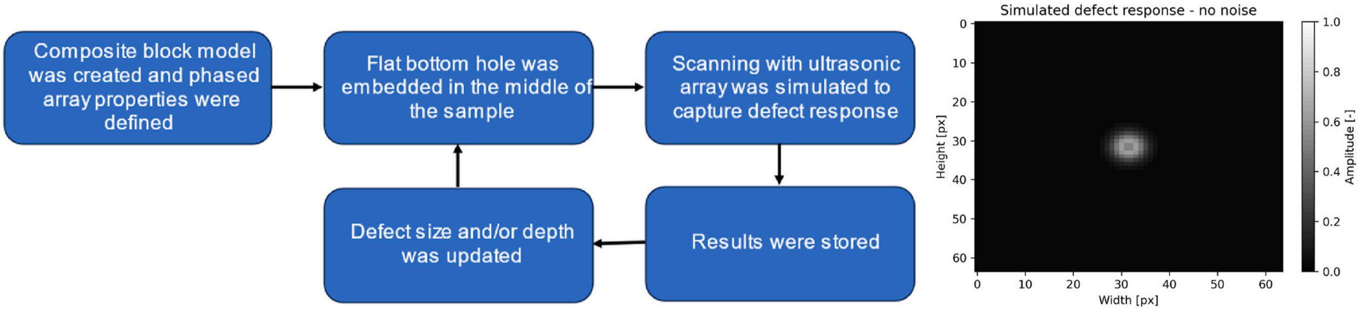

Composite block model was created and phased array properties were defined
Flat bottom hole was embedded in the middle of the sample
Scanning with ultrasonic array was simulated to capture defect response
Defect size and/or depth was updated
Results were stored
Simulated defect response - no noise
Height [px]
Width [px]
Amplitude [-]
0
10
20
30
40
50
60
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 5. Simulation process flow chart for the parametric sweep of defect dimension and location (left) and an example of simulated C-scan image of a 6.0 mm FBH at 4.5 mm depth (right).

thickness of each layer being 1 mm. This was different from the experimental samples which was made with non-crimp fabric. Fibre layers were considered transversely isotropic with a density of 1670 kg/m^3 while polymer matrix was defined as isotropic material with a density of 1230 kg/m^3. Longitudinal and transversal wave velocities were set at 2488 m/s and 1134 m/s respectively. These values were determined experimentally by conducting an ultrasonic scan on sample E with known thickness. Next, on ultrasonic data the distance between front and back wall reflections was calculated and correlated with the sampling rate of the ultrasonic controller. Lastly, the speed of sound was calculated with:

$$v = \frac{2 * d}{n_{samples}/f_s} \qquad \text{Eq (1)}$$

Where v is the speed of sound in m/s, d is the thickness of the material in m, and $n_{samples}$ is the number of samples in the ultrasonic data between the front and back wall responses, and f_s is the sampling frequency of the equipment which was set at 100 MHz. Wave attenuation was set to follow the power attenuation law given by Equation (2):

$$\alpha(f) = \sum_{p=1}^{n} \alpha_p * f^p \qquad \text{Eq (2)}$$

Where α_p is wave attenuation given in dB/mm, f is the frequency in Hz, and p is the power of the frequency. For this study α_p was set at 0.815 dB/mm and p was 4.

To create a scanning path simulation, an immersion linear phased array with 64 elements, 0.8 mm pitch and an element gap of 0.1 mm was modelled with a stand-off of 20 mm from the sample filled with water with no assumed attenuation and a velocity of 1483 m/s. The operating frequency of the array was set to 5 MHz with Hanning windowing and a 100 MHz sampling rate. Scanning was performed in linear mode with a sub-aperture of 4 elements to match the experimental setup. The step of array movement was set to match the array element pitch of 0.8 mm and was moved across the defect in a total of 64 steps. Subsequently, 3D volumetric data was stored as a 61 × 64 x 1000 array (Number of sub-apertures, steps of array scan, and spatial samples for each A-scan) and processed with the signal processing method described in section 2.3. Overall, this resulted in total of 300 simulations of FBHs that were used during the training of the ML models.

2.3. Signal processing and imaging

Simulated data and captured experimental data were stored as three-dimensional arrays (in format of [number of sub aperture, number of scan position, time data]) comprised of all A-scans collected along the electronic and mechanical scanning direction of the array and the robotic arm, respectively (refer to Fig. 4). Data were normalised with respect to the maximum amplitude occurring across all captured A-scans. Next, a Hilbert transform was applied to each A-scan to extract the envelope of the signal. This processing method is often used with UT in NDE to improve amplitude response by adding phase shifted signal to the original one [64]. Moreover, the resulting A-scans were time gated to remove the front and back wall responses. Time gating was done manually for each individual sample due to the varying material thicknesses. Lastly, maximum amplitudes of gated signals were used to construct a C-scan image. Examples of normalised A-scan, Hilbert A-scan, and amplitude C-scan are shown in Fig. 6.

2.4. Augmentation of synthetic data

When comparing the C-scans presented in Figs. 5 and 6, there is a clear difference in structural noise that CIVA model failed to capture. This also adversely affects the defect response as the defect indication from the CIVA model looks undisturbed and very uniform. ML models benefit from training on data that represents reality as accurately as possible; therefore, a novel post-processing approach should be devised to overcome the lack of modelling noise which is present in the experimental data. To this end, the method of A-scan noise addition proposed by Ref. [54] was implemented. This foundation of noise augmentation approach is because each A-scan is composed of structural noise, resulting from interactions between individual material layers, and random noise from sources such as electrical interference. The authors have demonstrated that this approach improves the performance of machine learning models compared to the use of raw simulated data. For noise profile analysis, pristine CFRP sample E was scanned with the experimental setup described in section 2.1.

To eliminate random noise introduced by external sources and extract structural noise, all A-scans in the complete scan were averaged. Next, all A-scans of a single B-scan were averaged and compared to the mean A-scan calculated in the first step. The difference between these two A-scans represents the structural noise component. This process was repeated for each B-scan in the scan. The resulting data was plotted in a histogram and described with normal distribution with a standard deviation of 0.003. For the random noise component, the averaged A-scans from each B-scan were subtracted from individual A-scan. This process was repeated for all B-scans. The resulting data was approximated with normal distribution with a standard deviation of 0.013. The process for the calculation of structural and random noise is presented in Fig. 7.

The generation of a new noise profile was performed by applying mean structural noise and adding a variance that corresponds to the approximately normal distribution. Following this, the random noise component is added with the mean and variances calculated in the previous steps.

Fig. 8 illustrates the simulated response, generated noise, and the final combined synthetic image. Note that the scale bar is modified to emphasize noise in the final image. Without scale bar changes it would be difficult to observe the additional noise in the final image due to relatively low amplitudes of added noise.

2.5. Amplitude image thresholding

The first method of defect detection that was explored in this work was amplitude image thresholding. In industry, a 6 dB drop on A-scans is often used for defect sizing, but in this work the approach is adapted for

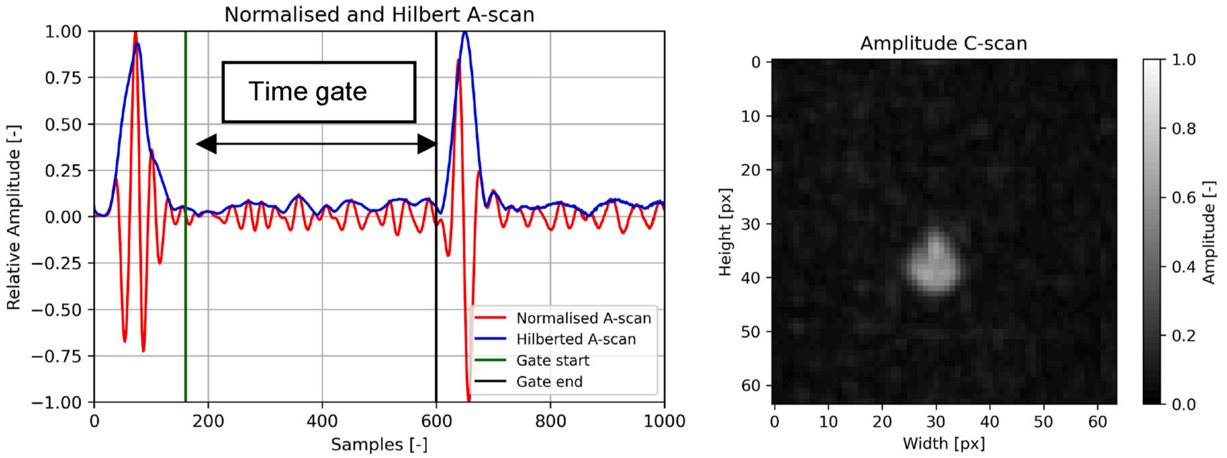


Fig. 6. Normalised A-scan and Hilbert processed A-scan with gating window (left) and an example of C-scan (right).

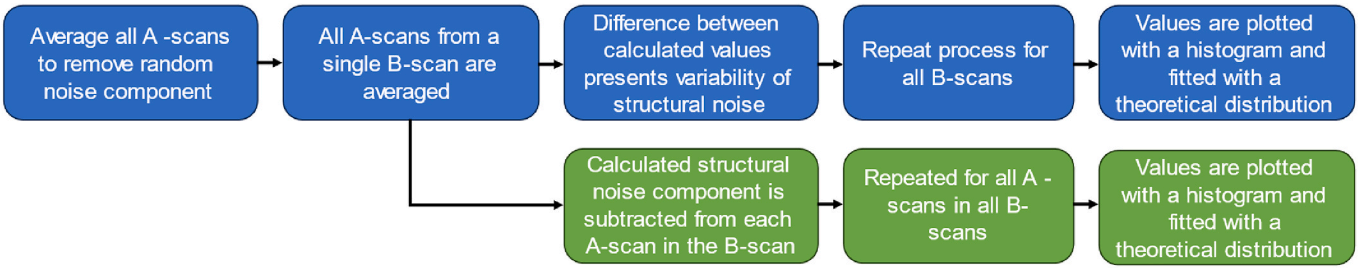

Average all A -scans to remove random noise component
All A-scans from a single B-scan are averaged
Difference between calculated values presents variability of structural noise
Repeat process for all B-scans
Values are plotted with a histogram and fitted with a theoretical distribution
Calculated structural noise component is subtracted from each A-scan in the B-scan
Repeated for all A - scans in all B-scans
Values are plotted with a histogram and fitted with a theoretical distribution

Fig. 7. Process for determination of structural (blue) and random (green) noise components (bottom). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

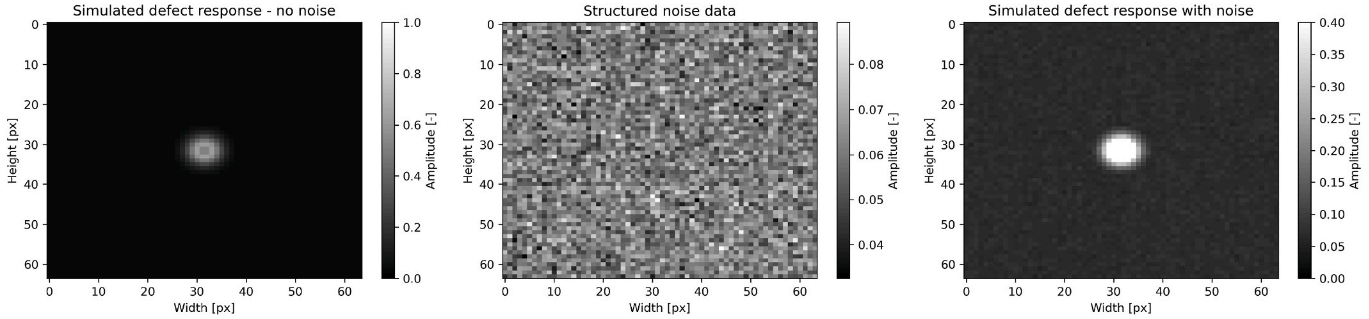


Fig. 8. Representation of augmentation results. Simulated response (left), generated noise (middle), and combined image (right).

defect detection and localization. Physically, the 6 dB drop on an A-scan signal represents the positions/time samples at which the maximum signal response losses half of its amplitude. In the ideal case, assuming A-scans were gated properly to exclude the front and back wall echoes, the maximum signal response would be created from the strong scatterers such as delaminations. Similarly, this loss of amplitude can be examined at an amplitude level other than half of the original value (e.g., 9, 12 or 18 dB). While used frequently, the 6 dB drop often performs poorly when it comes to larger defects of irregular shape and defects that are smaller than the beam width of the acoustic wave [45]. A smaller body of research tackled this issue and analysed alternatives to the 6 dB method [44,45,65]. The amplitude thresholding approach is hereby used as a basis for comparing the traditional methods to the ML networks tested in section 2.7.

To apply the amplitude thresholding method to the ultrasound experimental data, amplitude C-scans were created following the procedure described in section 2.3. Subsequently, the maximum pixel value of the resulting image was found, and the image was thresholded for 6-, 9-, and 12-dB drops (corresponding to 50%, 65% and 75% losses of amplitude). All pixels that had values lower than the calculated threshold were set to 0, while those with values larger than the threshold were set to 1, creating a binary map of the original image. Next, the spaghetti algorithm [66] was used to find connected components. The algorithm selects an unmarked pixel and assigns it to a new connected component, and afterwards it moves to neighbour pixels and assigns them to the same connected component. This process is repeated until all pixels are assigned. Furthermore, the algorithm produces coordinates and areas of connected components. Lastly, resulting coordinates are used to create rectangles that encapsulate the corresponding defective area. For display purposes, these rectangles were overlaid over the original image.

2.6. Statistical image thresholding

In addition to the previous method, a statistics-based approach was also evaluated. This process is based on work presented in Ref. [53] where no prior knowledge about defects is needed, only that they have sufficiently different acoustical responses than defect-free areas. Firstly, a representative defect-free section of the amplitude C-scan from sample E was extracted and used for statistical analysis. The goal of this method is to convert pixel values to probability values, where a higher number indicates a higher probability that an individual pixel belongs to a defect class. The pixel amplitudes in the extracted section were represented by a histogram, with a number of bins calculated with the Freedman-Diaconis rule [67]. Next, the SciPy Python package was used to test theoretical distributions and determine the best Probability Density Function (PDF). PDF is the mathematical representation that describes likelihood or probability of observing different values for some continuous variable. By extension, a Cumulative Density Function (CDF) is also computed. CDF is a related concept to PDF, as it indicates the probability of encountering the value that is less or equal to a point described by the PDF. Lastly, each pixel value from the original image was remapped to a corresponding probability according to the CDF. For current set of data, an f-distribution was determined to provide the best fit to the histogram. A range of probabilities was used (99, 99.5, 99.9%) to determine defective areas in the remapped image. An example of generated PDF and CDF for sample E is presented in Fig. 9.

2.7. Object detection neural networks

In this work, the defect detection performances of machine learning models from You Only Look Once, Faster R–CNN, and RetinaNet family of models were compared. Details for tested network are included in appendix A. The choice of networks stems from their track record as state-of-the-art models on various object detection datasets, and from variations in their architecture that influence their inference speed and performance. Furthermore, PyTorch library made pretrained weights for these models readily available, streamlining the process of transfer learning. To this end, all networks were pretrained on COCO datasets. Full fine-tuning was performed, and no layers were frozen, except for the final classification layer which was adjusted to conform to classes used in this work. Furthermore, all tested ML and thresholding methods were evaluated on the same images to ensure consistency. Due to the nature of Hilbert transformed ultrasonic C-scans, the data was bound between

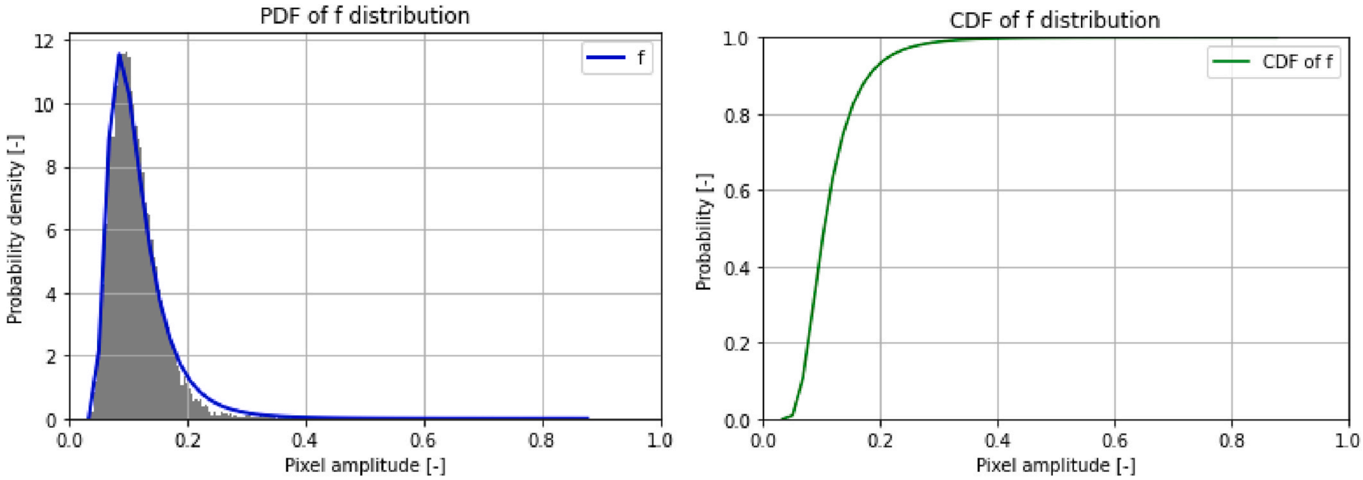


Fig. 9. Probability density function (left) and resulting cumulative density function (right) of sample E.

0 and 1.

Training dataset consisted of 300 synthetically generated C-scan images of size 64 × 64 pixels. The range of generated defects ranged from 3.0 mm to 15.0 mm, at 12 different depths starting at 1.5 mm measured from the front surface and extending to 7.0 mm. A subset of this training dataset (10%) was used as a validation dataset. The data splitting was performed randomly using fixed random seeds to ensure repeatability of the training process for all networks. All defects were circular in shape, with no deviations in shape or position within the generated image. A separate experimentally acquired testing dataset consisted of 8 amplitude C-scans, containing a range of defect types and sizes. FBHs were present in samples A and B, while other samples contained Teflon inserts and bagging films which were rectangular in shape. For more detail about testing dataset please refer to section 2.1. and Appendix B.

All proposed networks were trained using a desktop PC equipped with Nvidia RTX 3090 GPU, 128 GB RAM and Intel® Xeon® Gold 6428 2.50 GHz CPU. Windows 11 with PyTorch library and Python version 3.10.8 were used for both training and evaluation of the models. An overview of the training hyperparameters for all the models used for training are presented in Table 2. Hyperparameters were chosen according to values proposed by the original authors of the used models. It is worth noting that a more extensive hyperparameter optimisation and model modification could lead to positive improvements, however, this was not in the scope of the current research, and it will be pursued in the future work. During training, data augmentation in form of random image translation, scaling, vertical, and horizontal flipping was introduced, except for YOLO family of models, which additionally employed mosaic augmentation. During model deployment onto the test dataset, no augmentations were used. All models were trained for 50 epochs, and model weights at the lowest validation score were saved.

Table 2
Overview of used training hyperparameters.

Hyperparameter/ Model	Yolov5 - Medium	Yolov5 - Large	Faster R–CNN	RetinaNet
Epochs	50	50	50	50
Learning rate	0.01	0.01	0.005	0.0005
Momentum	0.937	0.937	0.9	0.9
Optimizer	SGD	SGD	SGD	SGD
Batch size	32	32	32	32
Weight decay	0.0005	0.0005	0.0005	0.0005
Model size in megabytes	42 MB	92 MB	159.7 MB	130.27 MB
Parameters	21.2 M	47 M	41.8 M	34.0 M
Backbone	CSP-Darknet53	CSP-Darknet53	ResNet50-FPN	ResNet50-FPN

2.7.1. *Performance metrics*

Each machine learning experiment consisted of training 10 networks using a set of random seeds to ensure the repeatability of experiments. In this study precision, recall, and F1 score were used as evaluation metrics. Precision is defined in Equation (3), and it illustrates the percentage of positive predictions that are correct according to the ground truth.

$$P = \frac{TP}{TP + FP} \qquad \text{Eq (3)}$$

In the equation above TP annotates true positives and FP annotates false positives. The recall is defined as the likelihood of detecting objects determined by ground truth. Mathematically this is represented in Equation (4):

$$R = \frac{TP}{TP + FN} \qquad \text{Eq (4)}$$

In the second equation, FN denotes false negatives. F1 score is defined as a harmonic mean of precision and recall and is shown in Equation 5:

$$F1 = 2 * \frac{P * R}{P + R} \qquad \text{Eq (5)}$$

For NDE applications, recall is more important as it is crucial to not miss any defects while having some false positives is tolerable at the expanse of adding to the analysis time. To evaluate which predictions are considered positive, Intersection Over Union (IoU) is used. IoU is represented in Equation 6:

$$IoU = \frac{Pred \cap GT}{Pred \cup GT} \qquad \text{Eq (5)}$$

Where *Pred* denotes a bounding box prediction, and *GT* denotes Ground Truth. Furthermore, for a complete view of the model performance, precision-recall curves were constructed and Area Under Curve (AUC) were reported. For this work it was decided to use a IoU value of 0.25, as for this application it is not as important to capture the full extent of the damage, and even the smaller predictions should be considered as positive results.

3. Results and discussion

In sample A, the application of amplitude thresholding with a 6 dB drop failed to identify four 3.0 mm FBHs and two smaller delaminations. This failure was attributed to the presence of stronger reflectors in the scan, specifically shallower 9.0 mm FBHs, which contained the maximum amplitude of the image. Similar observations were made in sample B containing FBHs, where a single 4.0 mm FBH and several 3.0

mm FBHs went undetected. Furthermore, in both samples C and D, the 6 dB method proved inadequate in identifying shallower and smaller indications, resulting in a poor overall defect detection performance with only 38.8% of the defects being correctly identified. Such low performance is also attributed to the defined IoU level of 0.25, as some predictions were made in the correct area but were much smaller than the provided ground truth.

The use of a more aggressive 9 dB drop method led to the identification of more defects. However, in the samples with FBHs, the shallowest 3.0 mm defect and two small delaminations were once again missed. The 9 dB drop method performed well in detecting all Teflon and bagging film inserts, however several false negative indications started to appear. This issue was particularly prominent in sample C, which exhibited brighter areas in the scan due to imperfections during the scanning process and the application of gating parameters for image creation. In this case, the gating process for C-scan generation incorporated some reverberations from the front wall, which were misinterpreted as defective areas. Compared to the 6 dB drop method, the 9 dB drop method achieved a much higher defect detection rate of 72.5%.

Lastly, the 12 dB drop method successfully identified most defects, albeit with an even higher number of false positive indications. This problem was again most pronounced in sample C with Teflon inserts, lowering the overall precision to 53.0%. In conclusion, amplitude thresholding of amplitude c-scans can yield satisfactory results for reflective defects when proper gating techniques are employed. However, this approach may face challenges when defects are located close to the samples' front surface, as the gating process may include front wall reverberations with high amplitudes. Additionally, this method proves unreliable in instances where no defects are present in the scan, as numerous areas are erroneously marked as defective due to the maximum amplitude being taken from structural noise. Lastly, even with IoU set at a relatively low threshold of 0.25, certain predictions are marked as false positives despite them correctly identifying a small area of the defect. With the increase of IoU, results of amplitude thresholding would deteriorate even further. When it comes to maximum achieved F1 score, 9 dB drop produced the best results at 70.3% F1.

With the statistical method, high probability values must be used to filter out false positive detections. In sample A, even though majority of defects were detected, the number of false positive indications outweighed the correct indications significantly when a 99% probability threshold was employed. The same trend was observed in samples C and D, where numerous false indications compromised the overall performance of the method with precision and F1 scores being only 50.8%. Despite this, a total of 95.0% of the defects were located successfully.

By increasing the probability to 99.5%, the number of false positives decreased. This adjustment had a positive impact on both the overall precision and F1 score, resulting in increases of 8.8% and 8.1%. With a recall rate of 91.2%, the statistical image thresholding method outperformed the amplitude threshold technique, but with lower precision. Furthermore, similar to the amplitude thresholding method, the statistical approach generated false positive indications when features other than defects with higher amplitudes were present in the image. The statistical method exhibited high sensitivity to gating parameters, which greatly influenced the number of false positive indications. Notably, when testing this method on pristine samples, no false detections occurred as the pixel values were close to the mean of the statistical distribution, without obvious outliers. Similar trend continued for the 99.9% threshold, where precision increased to 64.9%, but the recall dropped to 76.2%. Precision of statistical thresholding method could be improved by imposing an additional area threshold in the predictions, but this risks filtering of smaller defects. Overall, the presented method provides an improvement over the amplitude thresholding method, especially in the recall values, with room for improvement when it comes to its precision.

The Faster R–CNN implementation trained on raw data detected all larger defects, but it consistently struggled to detect the smallest and deepest FBH defects. On average, the Faster R–CNN model performed well in generalising to rectangular-shaped defects and FBHs that are 4 mm or larger.

The major benefit of this implementation is a greatly improved precision score of 98.6% when compared to the statistical method and amplitude thresholding methods. This improvement is attributed to the ability of machine learning models to learn complex features that describe defective areas, whereas previous methods relied solely on amplitude values. As a result, the robustness of the machine learning model matches that of previous methods while providing increased resilience to imperfections in the scanning process and signal gating. When data augmentation was performed, it resulted in minor increases in precision, recall, F1, and AUC score (1.1%, 1.2%, 1.1%, and 1.1% increase, respectively). Nevertheless, both the Faster R–CNN trained on raw data and the augmented data show improvements over previous techniques by providing a more robust detection mechanism, with significant enhancements in precision and F1 metrics.

Similar to the Faster R–CNN, RetinaNet provided an increase in precision when compared to statistical thresholding, but still oftentimes missed smaller 3.0 and 4.0 mm FBHs. RetinaNet generalised well on the rectangular-shaped defects in samples C and D but had several false indications that were very close to the positive indications. Such false indications were refined with a better choice of non-maximum suppression (NMS) thresholds. Furthermore, there were some instances where indications captured multiple defects under the same bounding boxes. These were considered false positives as clear separation between the defects is important, even when they are close in distance. Upon augmentation, precision dropped by 3.6%, but the recall rate increased by 4.0%. The main difference was that 3.0 and 4.0 mm defects were identified with greater recall rates than with the Faster R–CNN.

The medium YOLO model identified most of the defects, however it struggled with sample D where some rectangular defects were missed. This observation implies that these networks could generalise better to the rectangular defects if some examples are included in the training dataset. Interestingly, similarly sized defects were identified in other samples, indicating a potential discrepancy in aspect ratios between the data used during training and inference as a possible cause. Sample D was created using a single ultrasonic pass compared to 3–5 passes in other samples, which resulted in a more extreme aspect ratio of visualised data. Consequently, this produced a significantly smaller image, with the width of the scan considerably narrower than its height. The YOLO model is more susceptible to changes in aspect ratios due to its use of defined anchor boxes. Aligning the aspect ratios of training data with that of test data could mitigate this effect, potentially improving the model's performance in scenarios with varying aspect ratios. Furthermore, it was possible to detect the missed defects by lowering the confidence threshold during inference, but it resulted in a higher overall number of false positives. An overall AUC of 87.0% and a maximum F1 score of 91.5% was achieved. Augmentation of the data resulted in a minor increase in AUC and F1 scores, of 0.6% and 0.5% respectively. Recall remained the same, therefore an increase in precision positively impacted the F1 score.

The large YOLO models achieved similar results, all defects from samples A, B and C were identified. This was an interesting observation as a large YOLO network generalised to FBHs of all sizes even without augmentation. All missed detection came from sample D where networks struggled to detect smaller rectangular Teflon inserts. This type of defect was not present in the training data, which indicates that this network could benefit from the inclusion of some examples of rectangular defects. The addition of augmentation yielded improvements of 1.4% in AUC, 2.4% in precision, and same recall at 86.9%. It is worth noting that YOLO family of models produce more bounding boxes of lower confidences, and the results are heavily influenced by NMS and confidence thresholds.

Overall, all ML models provided an improvement over the statistical thresholding and amplitude thresholding methods, even when trained

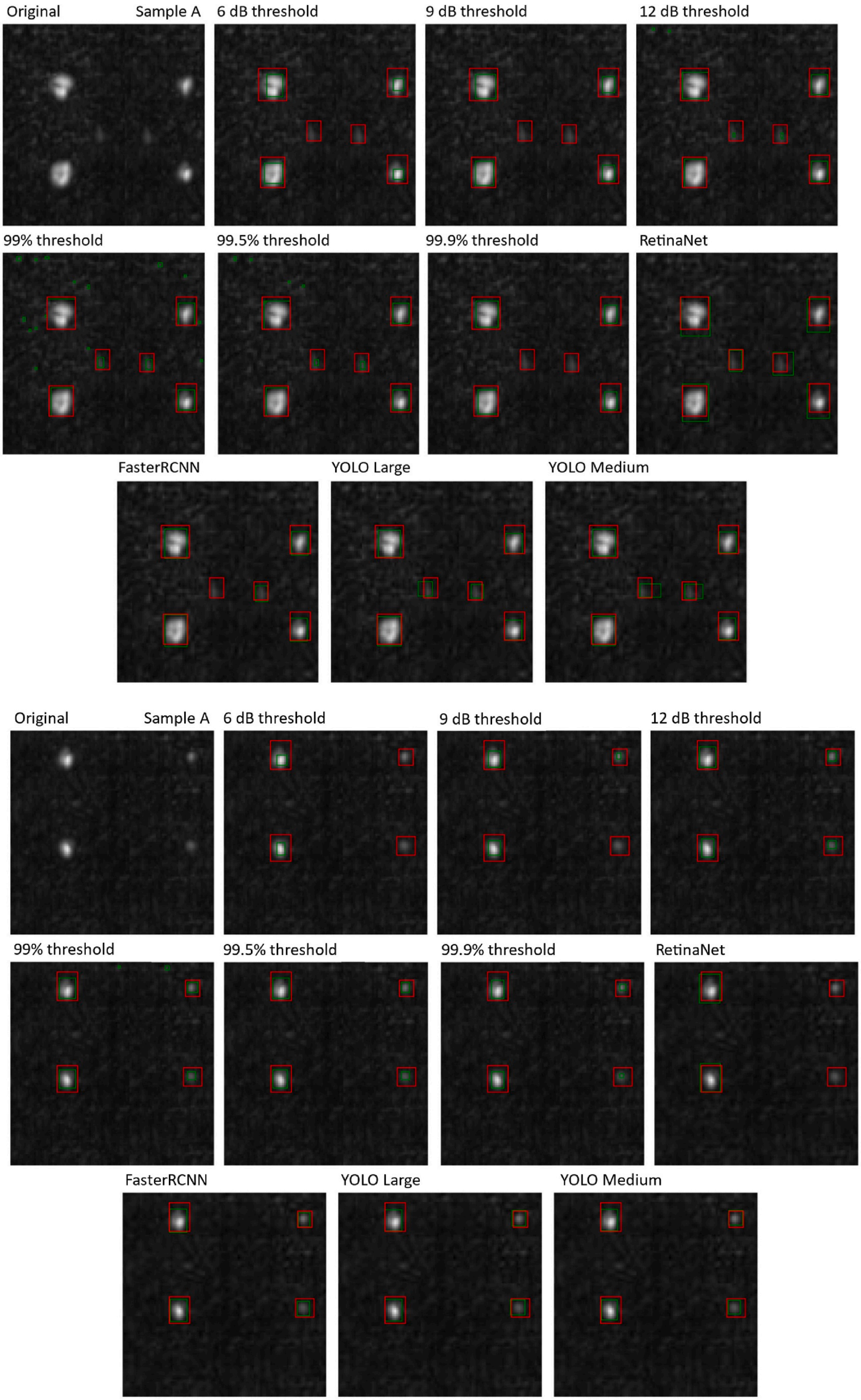


Fig. 10. Extracted sections of several testing images. Names of samples and used detection method are listed above each example, with ground truth bounding box marked in red and test results in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

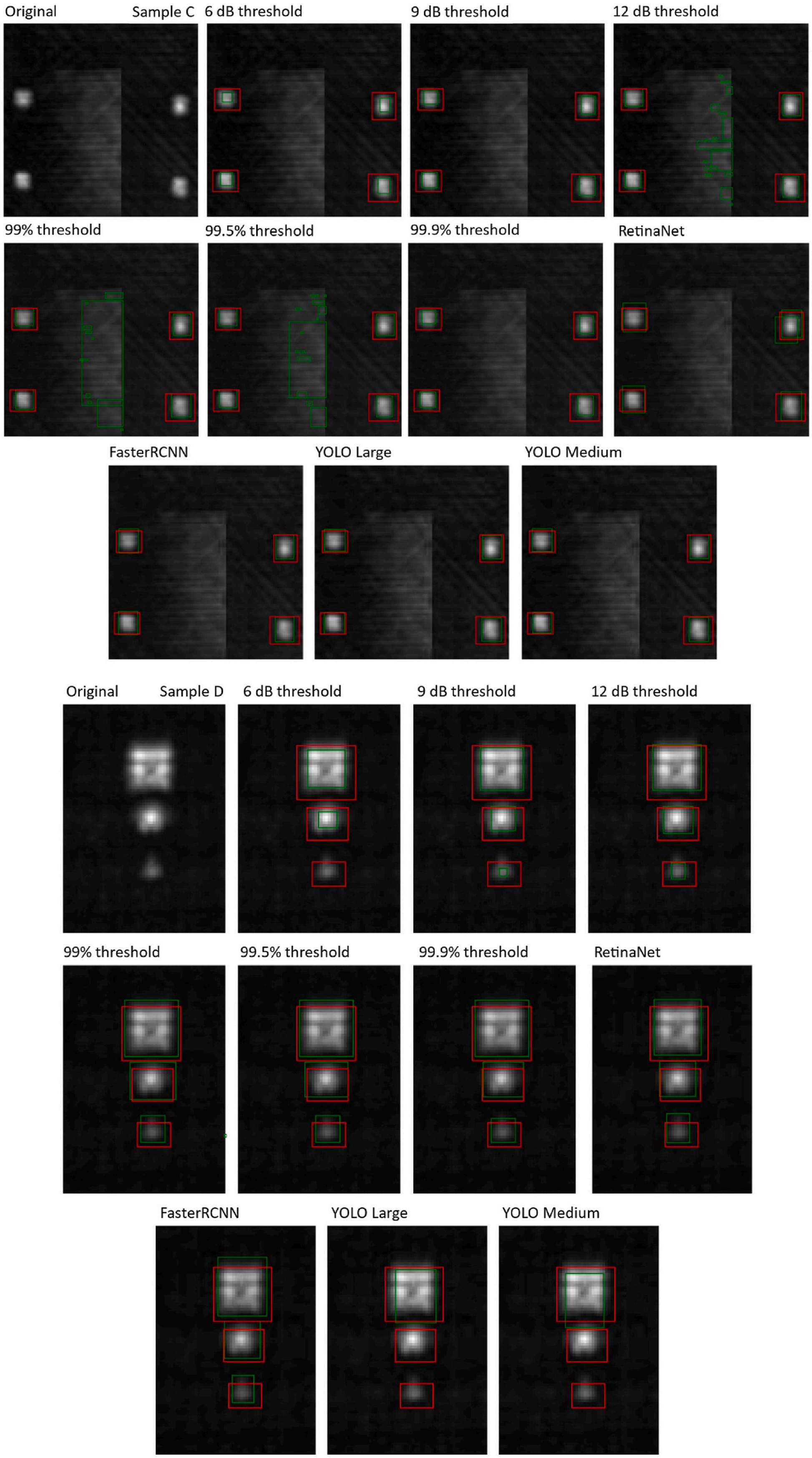


Fig. 10. (*continued*).

on raw simulation data. The augmentation of training data positively impacted recall and F1 scores of all models, with the minor exception of large YOLO model. The augmentation led to most prominent increases in FasterRCNN, which overall produced best results for this dataset. Furthermore, with the optimisation of confidence, IoU, and NMS thresholds, this score could be refined further. Another improvement would be the implementation of an ensemble of different machine learning models, where several models would process the same input and provide a combined output. The good results achieved in terms of recall with amplitude and statistical thresholding were always followed by a low score in precision and vice versa, making it hard to strike a balance between all performance metrics. On the other hand, machine learning provided more balanced and robust results throughout different tests. Visual representation of test results is illustrated in Fig. 10, with ground truth being presented with red bounding boxes and test results with green bounding boxes. Precision recall curves and evaluation metrics of all tested methods are presented in Fig. 11 and Table 3. Precision and recall scores in Table 3 are reported based on the maximum achieved F1 score.

Training times were modest due to the small dataset size, and the full computational times are presented in Table 4. Medium YOLO model was the fastest to train and the fastest ML model in the inference. Amplitude thresholding was the overall fastest method, mainly due to its simplicity. Statistical thresholding took 1250.8 ms per image, but this time is heavily influenced by the number of tested statistical distributions. The fitting process is repeated for each new image, but for scans with similar backscattering noise it is possible to perform this process only once and significantly speed up the inference times. In reported results, five candidate distributions were tested.

Table 3
Evaluation metrics for experimental dataset IoU = 0.25

Method	Training data/ Type	AUC	Precision	Recall	F1
Faster RCNN	Raw	0.961	0.986	0.949	0.967
	Augmented	0.972	0.998	0.960	0.978
RetinaNet	Raw	0.974	0.984	0.909	0.945
	Augmented	0.965	0.948	0.949	0.949
YOLO medium	Raw	0.870	0.979	0.859	0.915
	Augmented	0.876	0.992	0.859	0.920
YOLO large	Raw	0.878	0.985	0.869	0.923
	Augmented	0.892	0.982	0.869	0.922
Amplitude thresholding – 6 dB	N/A	N/A	0.456	0.388	0.419
Amplitude thresholding – 9 dB	N/A	N/A	0.682	0.725	0.703
Amplitude thresholding – 12 dB	N/A	N/A	0.530	0.887	0.664
Statistical thresholding – 99%	N/A	N/A	0.347	0.950	0.508
Statistical thresholding – 99.5%	N/A	N/A	0.435	0.912	0.589
Statistical thresholding – 99.9%	N/A	N/A	0.649	0.762	0.701

Table 4
Computational times.

Method	Training time [mins]	Testing/image [ms]
Faster RCNN	6.7	47.2
RetinaNet	11.4	79.8
YOLO Medium	2.4	44.9
YOLO Large	3.2	50.9
Amplitude Thresholding	N/A	0.3
Statistical Thresholding	N/A	1250.8 (1250.0 fitting + 0.81 thresholding)

4. Conclusion

In this paper, three different methods of defect detection and localization in the amplitude C-scans of CFRP samples were demonstrated: amplitude image thresholding, statistical image thresholding, and the use of object detection models. By mimicking the industrial NDE setup, a realistic data acquisition process with automated ultrasonic scanning on five different CFRP samples enabled the generation of representative datasets. The training of the ML models was driven by synthetic dataset generated by CIVA software, and further augmented by A-scan noise addition method, removing the need for use of experimental data in the training loop.

Through the investigations, it was concluded that:

- The amplitude thresholding method is suitable for the detection and localization of large reflective defects. However, this method was unable to detect smaller defects and was heavily reliant on the absence of any other reflective features that trigger false indications. Furthermore, this method performed poorly on scans where no defects were present.
- The improvement to this method is the statistical image thresholding method that outperforms amplitude thresholding in the sense that it

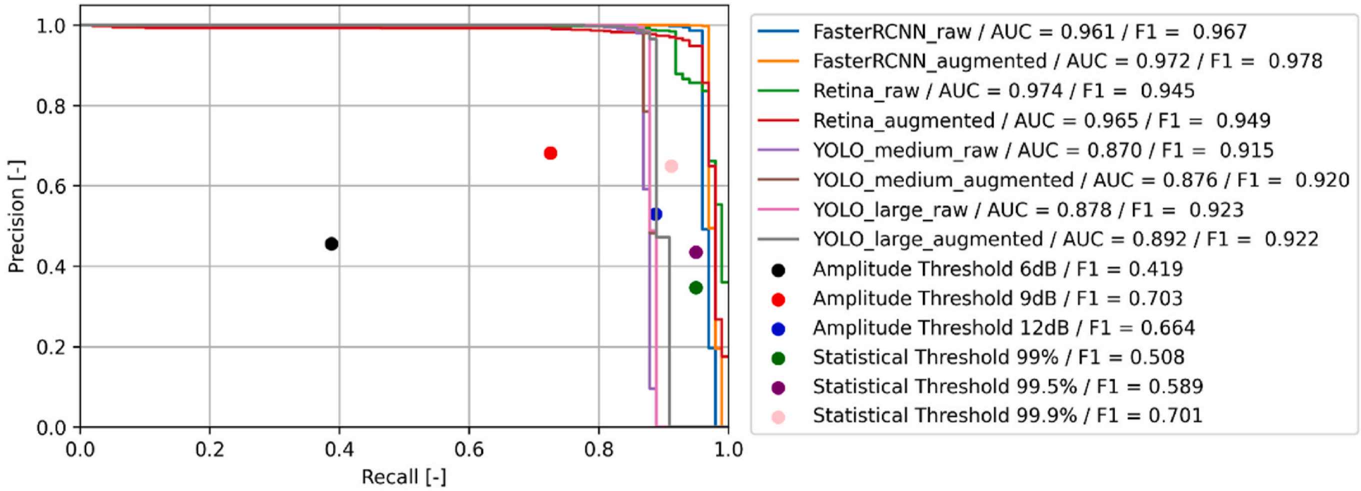


Fig. 11. Precision and recall curves for IoU = 0.25.

can process pristine samples without providing false indications. Unfortunately, the performance of this method was also reliant on the absence of bright artefacts in the images.

- Lastly, the four different ML algorithms tested for defect detection provided improvement over the statistical method, by accurately identifying defective areas. The performance of trained models was further improved by the application of a novel augmentation method. Validation of the models during the training process on a subset of the synthetic dataset was valuable as it diminished the need for the acquisition of a separate validation dataset.

For future work, the aim is to expand training and testing datasets to include different types of defects such as porosities and inclusions. While trained models demonstrate good results for detection and localization, the distinction between different types of defects was not achieved due to the limitations associated with available datasets. Furthermore, the model performance could be improved with more detailed hyperparameter optimisation and/or inclusion of ML ensembles in the process. One limitation of this work lies in the fact that performed scans were controlled and captured all defects in their entirety, and therefore no edge cases where only a portion of the defect was acquired. Exploration into the effect of varying the relative positioning of the array and defects is left to further work.

CRediT authorship contribution statement

Vedran Tunukovic: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Shaun McKnight:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing – review & editing. **Ehsan Mohseni:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing. **S. Gareth Pierce:** Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Richard Pyle:** Formal analysis, Investigation, Methodology, Visualization, Writing – review & editing. **Euan Duernberger:** Methodology, Software. **Charalampos Loukas:** Methodology, Software. **Randika K.W. Vithanage:** Supervision, Writing – review & editing. **David Lines:** Data curation, Methodology, Writing – review & editing. **Gordon Dobie:** Methodology, Supervision, Writing – review & editing. **Charles N. MacLeod:** Funding acquisition, Resources, Writing – review & editing. **Sandy Cochran:** Funding acquisition, Supervision, Writing – review & editing. **Tom O'Hare:** Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gareth Pierce reports financial support was provided by Spirit AeroSystems Inc.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported through EPSRC Centre for Doctoral Training in Future Ultrasonic Engineering (FUSE CDT) and Spirit AeroSystems/Royal Academy of Engineering Research Chair for In-Process Non-Destructive Testing of Composites, RCSRF 1920/10/32. Simulated data is available upon request.

Appendix A

Faster R–CNN

Faster R–CNN is a ML architecture released in 2015 as an improvement over the earlier R-CNNs models for object detection in images [37], introduced in Ref. [38]. The authors recognized that the region proposal step in R–CNN was the main bottleneck in terms of computational time and to address this they have introduced Region Proposal Networks (RPNs) to generate region proposals more efficiently. Faster R–CNN comprises of two components, RPNs and Fast R–CNN detector that performs object classification on the areas proposed by RPN. These two structures share convolutional layers which enable end-to-end training.

The RPN operates on the feature maps produced by earlier convolutional layers. It uses a sliding window approach, where a 3×3 kernel is moved over the feature map. At each position, the RPN predicts a set of region proposals, which are potential bounding boxes that might contain objects. These proposals are generated based on anchor boxes that have different scales and aspect ratios. Anchor boxes are predefined bounding box shapes that serve as a reference for generating region proposals. Following the RPN, the Fast R–CNN detector takes proposed regions and performs feature extraction using pooling layers that convert variable-sized regions proposals into fixed-size outputs. These outputs are then propagated through fully connected layers that perform classification. For training, the authors used a multi-task loss function which combines classification and bounding box regression losses. ResNet50-FPN, a variation on ResNet architecture introduced in Ref. [68], was used for feature extraction and creation of feature maps. Recommended hyperparameters used by the original authors were deployed, with changes to the batch size and training epochs. For robustness and faster training convergence, initial pre-trained weights from a Faster R–CNN model that was trained on the Microsoft (MS) COCO dataset [69] were adopted.

You only look once

You Only Look Once (YOLO) object detection models were initially developed by Redmon et al. [70], with multiple iterations being released in recent years from various research teams [41–43,71]. Compared to the region proposal and sliding windows methods used in R–CNN and Fast R–CNN, YOLO introduced techniques that improved both accuracy and inference speed. These include single-stage detection where both bounding box coordinates and class is determined with a single pass through the network and mosaic augmentation which enhances the training datasets. In this study, the implementation by the company Ultralytics [71] was utilised. This implementation includes architectures of varying sizes and complexities that were pre-trained on the MS COCO dataset [69]. All architecture variants share the common underlying structure consisting of a Cross Stage Partial (CSP) network in the backbone, a Path Aggregated Network (PAN) in the neck, and a YOLO v3 detection head.

The CSP network [72] was implemented in the backbone due to its efficiency and the ability to deploy trained models to setups with weaker CPUs and GPUs. CSP is based on DenseNet and introduces the splitting of the gradient flow, which greatly increases speed and performance (an unedited feature map is combined with a feature map that propagated through the dense layer). The focal point of CSP gradient splitting is convolutional layer with 1 × 1 kernel size, which is computationally efficient and used to increase the complexity of the architecture. The Spatial Pyramid Pooling (SPP) [73] layer block is located at the end of the backbone. SPP allows YOLO networks to accept input images of any resolution by max pooling of the same input multiple times with different kernel sizes and strides before concatenating them. With this method, the output is always of the same dimension, making it compatible for use in the subsequent layers. The neck is the central part of the YOLO structure, which comprises a series of network layers that collect and integrate various characteristics obtained from the backbone prior to passing them to the final detection layers. Neck of the YOLO model is the PAN, developed in 2018 [74] and first introduced in YOLOv4 [43]. In short, it is an improvement over the previous Feature Pyramid Network (FPN), which is based on feature maps of varying sizes. The improvement came from additional lateral connections between low- and high-level feature maps in the feature pyramid. Lastly, the head portion of the network produces predictions in the form of a vector with the class of the object and coordinates for the proposed bounding box.

RetinaNet

In 2017, Lin et al. [75] developed a single-stage object detection model called RetinaNet that achieved better scores than its two-stage counterparts. The novelty in this work is a new loss function that addresses the issues of class imbalances that can happen if cross entropy is used as a loss function during training. The new loss function is called "Focal loss" and it diminishes the losses by an order of one magnitude for high-probability examples like pristine CFRPs, while still retaining high losses for low-probability examples such as defects.

Like YOLO, RetinaNet is a one-stage object detector that uses an FPN network for multi-scale feature representation. The classification and bounding box regression are handled by two smaller task-specific neural networks. The new loss function was combined with ResNet-101 and Feature Pyramid Network (FPN) to create RetinaNet, a model that achieved state-of-the-art on the COCO dataset. However, with an inference time of 200 ms, the final performance was not suitable for real-time tasks. In this implementation, ResNet50-FPN was used as the backbone. Hyperparameters used by the original authors were followed, with changes to training epochs, confidence, and NMS thresholds. Similar to Faster R–CNN, pre-trained weights and biases were used.

Appendix B

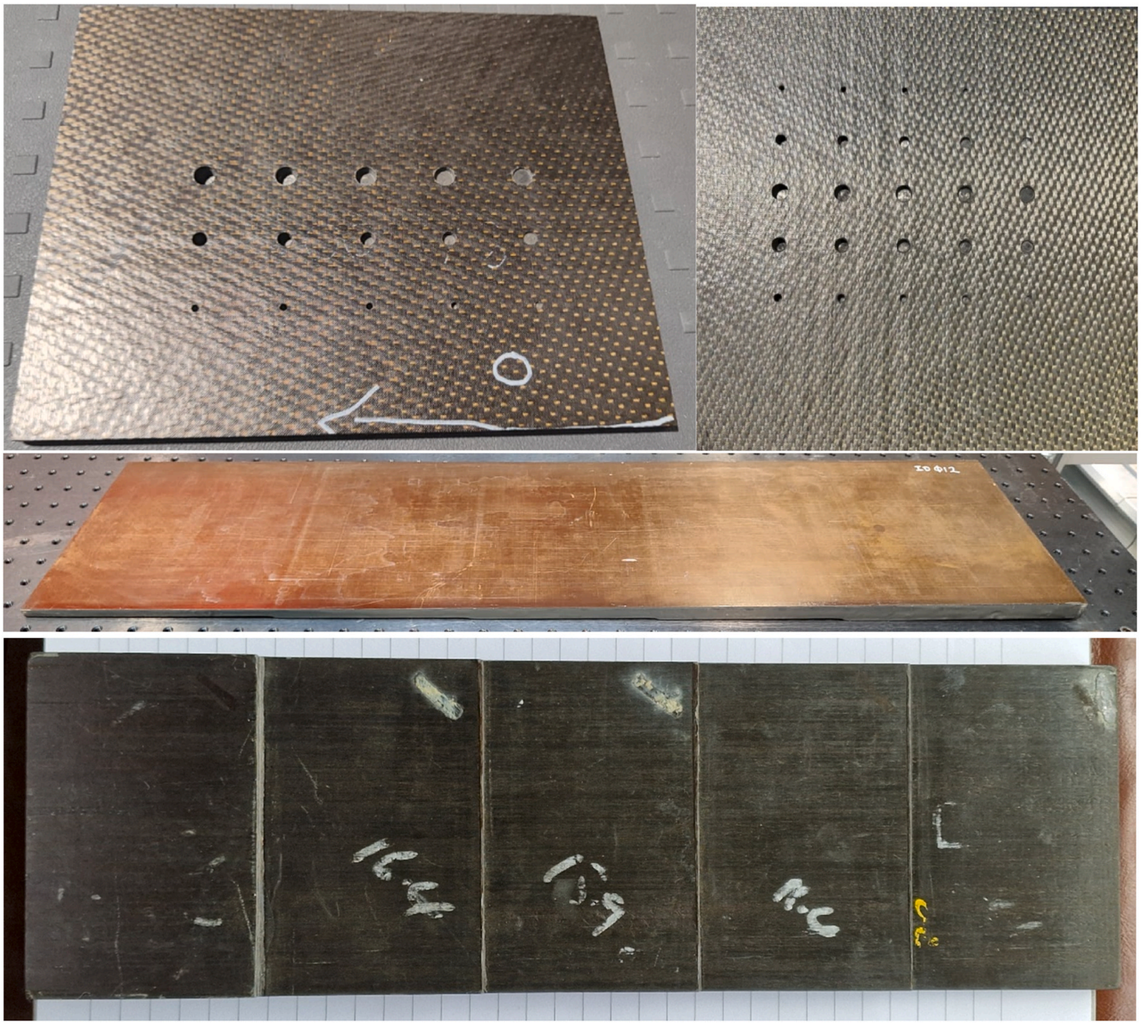


Fig. 12. Defective samples used in this study: Sample A (top left), Sample B (top right), Sample C (middle), and sample D (bottom)

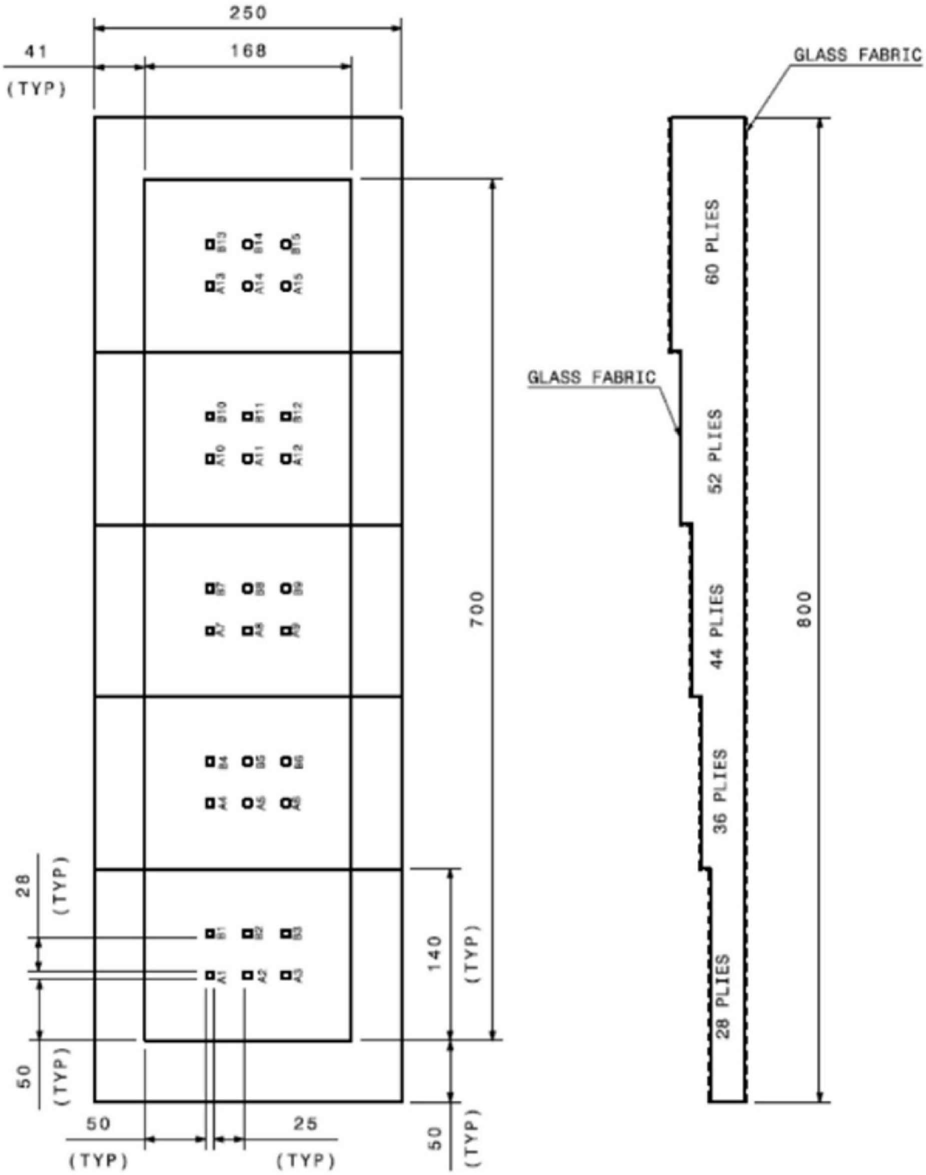


Fig. 13. CAD file for Sample C

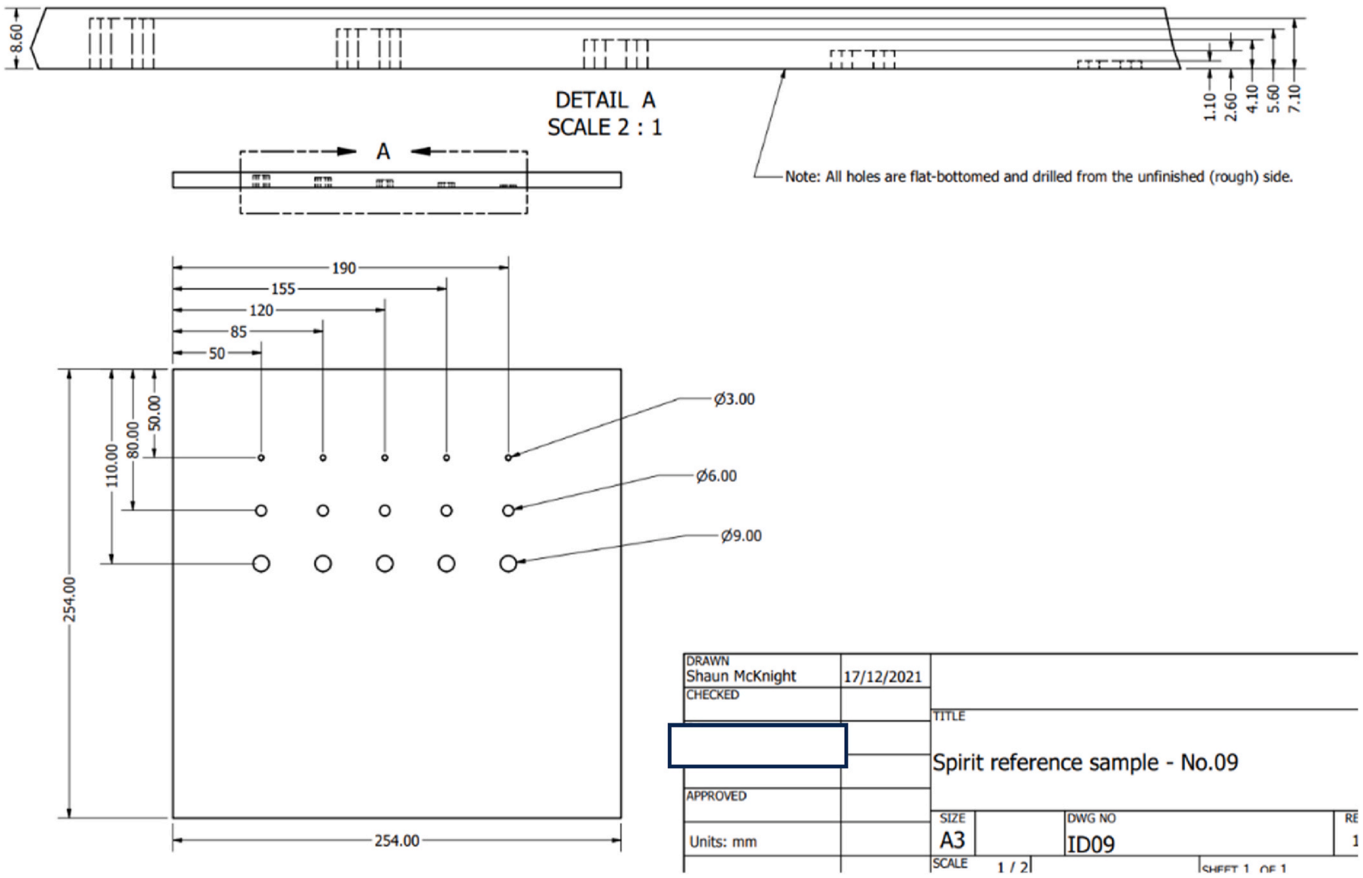


Fig. 14. CAD file for sample A

References

[1] Mangalgiri PD. Composite materials for aerospace applications. Bull Mater Sci 1999;22(3):657–64.

[2] Quilter A. Composites in aerospace applications. 2001.

[3] Slayton R, Spinardi G. Radical innovation in scaling up: Boeing's Dreamliner and the challenge of socio-technical transitions. Technovation Jan. 2016;47:47–58. https://doi.org/10.1016/j.technovation.2015.08.004.

[4] Giurgiutiu V. Structural health monitoring of aerospace composites. Elsevier; 2016. https://doi.org/10.1016/B978-0-12-409605-9.00001-5.

[5] Bachmann J, Hidalgo C, Bricout S. Environmental analysis of innovative sustainable composites with potential use in aviation sector—a life cycle assessment review. Sci China Technol Sci Sep. 2017;60(9):1301–17. https://doi.org/10.1007/S11431-016-9094-Y.

[6] Djordjevic B. Nondestructive test technology for the composites. In: The 10th international conference of the slovenian society for non-destructive testing; 2009. p. 259–65.

[7] Dragan K, Swiderski W. Studying efficiency of NDE techniques applied to composite materials in aerospace applications. Acta Phys Pol, A 2010:878–83.

[8] Ley O, Godinez V. Non-destructive evaluation (NDE) of aerospace composites: application of infrared (IR) thermography. In: Non-destructive evaluation (NDE) of polymer matrix composites: techniques and applications. Elsevier Ltd; 2013. p. 309–34. https://doi.org/10.1533/9780857093554.3.309.

[9] Schnars U, Henrich R. Applications of NDT methods on composite structures in aerospace industry. In: Conference on damage in composite materials; 2006.

[10] Kapadia A. National composites network best practice guide non destructive testing of composite materials. 2007 [Online], http://www.twi.co.uk/j32k/index.xtp. [Accessed 18 May 2022].

[11] Wooh SC, Shi Y. Optimum beam steering of linear phased arrays. Wave Motion 1999;29(3):245–65. https://doi.org/10.1016/S0165-2125(98)00039-0.

[12] 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 Dec. 2005;38(8):701–11. https://doi.org/10.1016/J.NDTEINT.2005.04.002.

[13] Wilcox PD. Ultrasonic arrays in NDE: beyond the B-scan. AIP Conf Proc Jan. 2013; 1511(1):33. https://doi.org/10.1063/1.4789029.

[14] Mineo C, et al. Flexible integration of robotics, ultrasonics and metrology for the inspection of aerospace components. In: AIP conf proc, vol. 1806; Feb. 2017, 020026. https://doi.org/10.1063/1.4974567. 1.

[15] Bertovic M, Virkkunen I. NDE 4.0: new paradigm for the NDE inspection personnel. Handb Nondestructive Eval 2021;4(0):1–31. https://doi.org/10.1007/978-3-030-48200-8_9-1.

[16] Virkkunen I, Koskinen T, Jessen-Juhler O, Rinta-aho J. Augmented ultrasonic data for machine learning. J Nondestr Eval Mar. 2021;40(1):1–11. https://doi.org/10.1007/S10921-020-00739-5/TABLES/1.

[17] Amiri N, Farrahi GH, Kashyzadeh KR, Chizari M. Applications of ultrasonic testing and machine learning methods to predict the static & fatigue behavior of spot-welded joints. J Manuf Process Apr. 2020;52:26–34. https://doi.org/10.1016/J.JMAPRO.2020.01.047.

[18] Ghafarallahi E, Farrahi GH, Amiri N. Acoustic simulation of ultrasonic testing and neural network used for diameter prediction of three-sheet spot welded joints. J Manuf Process Apr. 2021;64:1507–16. https://doi.org/10.1016/J.JMAPRO.2021.03.012.

[19] Siljama O, Koskinen T, Jessen-Juhler O, Virkkunen I. Automated flaw detection in multi-channel phased array ultrasonic data using machine learning. J Nondestr Eval Sep. 2021;40(3):1–13. https://doi.org/10.1007/S10921-021-00796-4/FIGURES/6.

[20] Koskinen T, Virkkunen I, Siljama Oskar, Jessen-Juhler Oskari. 'The effect of different flaw data to machine learning powered ultrasonic inspection'. J Nondestr Eval 2021;40:24. https://doi.org/10.1007/s10921-021-00757-x.

[21] Cruz FC, Simas Filho EF, Albuquerque MCS, Silva IC, Farias CTT, Gouvêa LL. Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing. Ultrasonics Jan. 2017;73:1–8. https://doi.org/10.1016/J.ULTRAS.2016.08.017.

[22] Munir N, Kim HJ, Park J, Song SJ, Kang SS. Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions. Ultrasonics Apr. 2019; 94:74–81. https://doi.org/10.1016/J.ULTRAS.2018.12.001.

[23] Munir N, Park J, Kim HJ, Song SJ, Kang SS. Performance enhancement of convolutional neural network for ultrasonic flaw classification by adopting autoencoder. NDT E Int Apr. 2020;111:102218. https://doi.org/10.1016/J.NDTEINT.2020.102218.

[24] Munir N, Kim H-J, Song S-J, Kang S-S. Investigation of deep neural network with drop out for ultrasonic flaw classification in weldments. J Mech Sci Technol 2018; 32(7):3073–80. https://doi.org/10.1007/s12206-018-0610-1.

[25] Medak D, Posilović L, Subašić M, Budimir M, Lončarić S. DefectDet: a deep learning architecture for detection of defects with extreme aspect ratios in ultrasonic

images. Neurocomputing Feb. 2022;473:107–15. https://doi.org/10.1016/J.NEUCOM.2021.12.008.

[26] Posilović L, Medak D, Subašić M, Budimir M, Lončarić S. Generating ultrasonic images indistinguishable from real images using Generative Adversarial Networks. Ultrasonics Feb. 2022;119:106610. https://doi.org/10.1016/J.ULTRAS.2021.106610.

[27] Medak D, Posilovic L, Subasic M, Budimir M, Loncaric S. Deep learning-based defect detection from sequences of ultrasonic B-scans. IEEE Sensor J Feb. 2022;22(3):2456–63. https://doi.org/10.1109/JSEN.2021.3134452.

[28] Medak D, Posilovic L, Subasic M, Budimir M, Loncaric S. Automated defect detection from ultrasonic images using deep learning. IEEE Trans Ultrason Ferroelectrics Freq Control Oct. 2021;68(10):3126–34. https://doi.org/10.1109/TUFFC.2021.3081750.

[29] Posilovic L, Medak D, Subasic M, Petkovic T, Budimir M, Loncaric S. Flaw detection from ultrasonic images using YOLO and SSD. Int Symposium Image Signal Process Anal ISPA Sep. 2019;2019-September:163–8. https://doi.org/10.1109/ISPA.2019.8868929.

[30] Zacharis P, West G, Dobie G, Lardner T, Gachagan Anthony. Data-driven analysis of ultrasonic inspection data of pressure tubes. Nucl Technol 2018;202(3):153–60. https://doi.org/10.1080/00295450.2017.1421803.

[31] Niu S, Srivastava V. Simulation trained CNN for accurate embedded crack length, location, and orientation prediction from ultrasound measurements. Int J Solid Struct May 2022;242. https://doi.org/10.1016/J.IJSOLSTR.2022.111521.

[32] Meng M, Chua YJ, Wouterson E, Ong CPK. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. Neurocomputing Sep. 2017;257:128–35. https://doi.org/10.1016/J.NEUCOM.2016.11.066.

[33] Guo Y, et al. Fully convolutional neural network with GRU for 3D braided composite material flaw detection. IEEE Access 2019;7:151180–8. https://doi.org/10.1109/ACCESS.2019.2946447.

[34] Tao C, Zhang C, Ji H, Qiu J. Fatigue damage characterization for composite laminates using deep learning and laser ultrasonic. Compos B Eng 2021;216(Jul). https://doi.org/10.1016/J.COMPOSITESB.2021.108816.

[35] Aldrin JC, Forsyth DS. Demonstration of using signal feature extraction and deep learning neural networks with ultrasonic data for detecting challenging discontinuities in composite panels. AIP Conf Proc May 2019;2102:230004. https://doi.org/10.1063/1.5099716/FORMAT/PDF.

[36] Nerlikar V, Mesnil O, Miorelli R, D'Almeida O. 'Damage detection with ultrasonic guided waves using machine learning and aggregated baselines'. Struct Health Monit 2023. https://doi.org/10.1177/14759217231169719/FORMAT/EPUB.

[37] Girshick R. Fast R-CNN 2015. Accessed: Jul. 14, 2022. [Online]. Available: https://github.com/rbgirshick/.

[38] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell Jun. 2015;39(6):1137–49. https://doi.org/10.48550/arxiv.1506.01497.

[39] Tan M, Pang R, Le Qv. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; Nov. 2019. p. 10778–87. https://doi.org/10.48550/arxiv.1911.09070.

[40] Tan M, Pang R, Le Qv. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; Nov. 2019. p. 10778–87. https://doi.org/10.48550/arxiv.1911.09070.

[41] Redmon J, Farhadi A. YOLOv3: an incremental improvement. Apr. 2018. https://doi.org/10.48550/arxiv.1804.02767.

[42] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017, 2017-January; Dec. 2016. p. 6517–25. https://doi.org/10.48550/arxiv.1612.08242.

[43] Bochkovskiy A, Wang C-Y, Liao H-YM. YOLOv4: optimal speed and accuracy of object detection. Apr. 2020. https://doi.org/10.48550/arxiv.2004.10934.

[44] Hauffe A, Hähnel F, Wolf K. Comparison of algorithms to quantify the damaged area in CFRP ultrasonic scans. Compos Struct Mar. 2020;235:111791. https://doi.org/10.1016/J.COMPSTRUCT.2019.111791.

[45] Li X, Wang Y, Ni P, Hu H, Song Y. Flaw sizing using ultrasonic C-scan imaging with dynamic thresholds. Insight: Non-Destructive Testing and Condition Monitoring Nov. 2017;59(11):603–8. https://doi.org/10.1784/INSI.2017.59.11.603.

[46] Barut S, Bissauge V, Ithurralde G, Claassens W. 'Computer-aided analysis of ultrasound data to speed-up the release of aerospace CFRP components', in *18th World Conference on Nondestructive Testing*, Durban, South Africa. e-Journal of Nondestructive Testing Apr. 2012;17(7).

[47] Song Y, Turner JA, Peng Z, Chen C, Li X. Enhanced ultrasonic flaw detection using an ultrahigh gain and time-dependent threshold. IEEE Trans Ultrason Ferroelectrics Freq Control Jul. 2018;65(7):1214–25. https://doi.org/10.1109/TUFFC.2018.2827464.

[48] Dogandžić A, Eua-Anant N, Dogandžic AD. Defect detection in correlated noise. AIP Conf Proc Apr. 2004;700(1):628. https://doi.org/10.1063/1.1711680.

[49] Wronkowicz A, Katunin A, Dragan K. Ultrasonic C-scan image processing using multilevel thresholding for damage evaluation in aircraft vertical stabilizer. Int J Image Graph Signal Process 2015.

[50] de Oliveira BCF, Nienheysen P, Baldo CR, Gonçalves AA, Schmitt RH. Improved impact damage characterisation in CFRP samples using the fusion of optical lock-in thermography and optical square-pulse shearography images. NDT E Int Apr. 2020;111:102215. https://doi.org/10.1016/J.NDTEINT.2020.102215.

[51] Osman A, Kaftandjian V, Hassler U, Rehak M, Hanke R. Steps toward automated 3D evaluation of ultrasound data. 2010. Accessed: Mar. 07, 2023, https://www.researchgate.net/publication/268296913.

[52] Li C, et al. Intelligent damage recognition of composite materials based on deep learning and ultrasonic testing. AIP Adv Dec. 2021;11(12):125227. https://doi.org/10.1063/5.0063615.

[53] Wilcox PD, et al. Fusion of multi-view ultrasonic data for increased detection performance in non-destructive evaluation. Proc R Soc A Nov. 2020;476(2243). https://doi.org/10.1098/RSPA.2020.0086.

[54] McKnight S, et al. GANs and alternative methods of synthetic noise generation for domain adaption of defect classification of Non-destructive ultrasonic testing. Jun. 2023. Accessed: Jun. 20, 2023, https://arxiv.org/abs/2306.01469v1.

[55] Blain P, et al. Artificial defects in CFRP composite structure for thermography and shearography nondestructive inspection Jun. 2017;10449(13):562–71. https://doi.org/10.1117/12.2271701.

[56] Vasilev M, et al. Sensor-enabled multi-robot system for automated welding and in-process ultrasonic NDE. Sensors Jul. 2021;21(15):5077. https://doi.org/10.3390/S21155077. 2021, Vol. 21, Page 5077.

[57] Robotics KUKA. KUKA KR90 R3100 extra HA specification manual. 2023. Accessed: Mar. 08, 2023. [Online], https://www.kuka.com/-/media/kuka-downloads/imported/8350ff3ca11642998dbdc81dcc2ed44c/0000208694_en.pdf.

[58] Olympus-ims 'RollerFORM. Phased array wheel probe manual'. 2023. https://www.olympus-ims.com/en/rollerform/. [Accessed 8 March 2023].

[59] Schunk. SCHUNK Force Torque sensors manual. 2023. https://schunk.com/us/en/automation-technology/force/torque-sensors/ft/ftn-gamma-si-130-10/p/EPIM_ID-30865. [Accessed 8 March 2023].

[60] MicoPulse 6PA | Phased Array Ultrasonic Technology | Peak NDT'. https://www.peakndt.com/products/micropulse-6pa/(accessed March. 8, 2023).

[61] Zhang Z, Liu M, Li Q, Ang Y. Visualized characterization of diversified defects in thick aerospace composites using ultrasonic B-scan. Compos Commun Dec. 2020;22:100435. https://doi.org/10.1016/J.COCO.2020.100435.

[62] 'EXTENDE, Experts in Non Destructive Testing Simulation with CIVA Software'. https://www.extende.com/(accessed December 26, 2022).

[63] Huthwaite P. Accelerated finite element elastodynamic simulations using the GPU. J Comput Phys Jan. 2014;257:687–707. https://doi.org/10.1016/J.JCP.2013.10.017.

[64] Drai R, Sellidj F, Khelil M, Benchaala A. Elaboration of some signal processing algorithms in ultrasonic techniques: application to materials NDT. Ultrasonics 2000;38:503–7. Accessed: Dec. 26, 2022, www.elsevier.nl/locate/ultras.

[65] Grosse CU, et al. Comparison of NDT techniques to evaluate CFRP-results obtained in a MAIzfp round robin test. 2016. Accessed: Jan. 30, 2023. [Online], http://creativecommons.org/licenses/by-nd/3.0/.

[66] F. Bolelli, S. Allegretti, L. Baraldi, and C. Grana, 'Spaghetti labeling: directed acyclic graphs for block-based connected components labeling'.

[67] Freedman D, Diaconis P. On the histogram as a density estimator:L2 theory. Z Wahrscheinlichkeitstheor Verwandte Geb Dec. 1981;57(4):453–76. https://doi.org/10.1007/BF01025868/METRICS.

[68] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Dec. 2016. https://doi.org/10.48550/arxiv.1612.03144.

[69] Lin TY, et al. Microsoft COCO: common objects in context. Lect Notes Comput Sci May 2014;8693(PART 5):740–55. https://doi.org/10.48550/arxiv.1405.0312. LNCS.

[70] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2016-December; Jun. 2015. p. 779–88. https://doi.org/10.48550/arxiv.1506.02640.

[71] Jocher G, et al. 'YOLOv5 SOTA realtime instance segmentation'. Nov. 22, 2022. https://zenodo.org/record/7347926. [Accessed 22 December 2022].

[72] Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW, Yeh IH. CSPNet: a new backbone that can enhance learning capability of CNN. In: IEEE computer society conference on computer vision and pattern recognition workshops, 2020-June; Nov. 2019. p. 1571–80. https://doi.org/10.48550/arxiv.1911.11929.

[73] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. Lect Notes Comput Sci Jun. 2014;8691(PART 3):346–61. https://doi.org/10.1007/978-3-319-10578-9_23. LNCS.

[74] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; Mar. 2018. p. 8759–68. https://doi.org/10.48550/arxiv.1803.01534.

[75] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, 2017-October; Dec. 2017. p. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.