University of Strathclyde Engineering

Royal Charter since 1964
Useful Learning since 1796

Deep Learning for Defect Analysis in Ultrasonic Non-Destructive Evaluation of Aerospace Composites

Shaun McKnight

Department of Electronic and Electrical Engineering

University of Strathclyde

A thesis submitted for the degree of

Doctor of Philosophy

August 2024

Copyright

Signed:

Date: 8th August 2024

Acknowledgements

In memory of "Pat" Bingham and "Bobby" McKnight.

This work would not have been possible without the support and contributions of many people. I firstly extend my heartfelt thanks to my supervision team: Dr Ehsan Mohseni, Prof. Gareth Pierce, and Prof. Charles MacLeod. Your guidance and mentorship have been invaluable. I was incredibly fortunate to have an experienced supervisory team that not only guided my work but also granted me the creative freedom to explore different ideas.

I am grateful to all members of the SEARCH lab, past and present, who have given their time and helped facilitate this work. I would like to thank my friends and colleagues, particularly Dr. Euan Duernberger and Dr. Euan Foster, for their exceptional technical support and friendship, which helped me settle smoothly into the start of my PhD project. It has also been a privilege to work closely with Vedran Tunukovic, sharing many of the highs and lows of research.

I would like to thank my industrial sponsor Spirit AeroSystems and the Royal Academy of Engineering Research Chair for funding this work. Along with the wider Spirit team, I would specifically like to thank Tom O'Hare for being a key contact for any industrial queries.

Finally, I wish to express my appreciation to my parents, Joan and John, whose love and support have been immeasurable. I am forever grateful for the opportunities you have afforded me.

Abstract

Carbon fibre reinforced polymer (CFRP) composites are critical materials in aerospace components and are seeing ever increasing demand. Ensuring the integrity of these composites is critical and Non-Destructive Evaluation (NDE) is widely used to provide crucial insights about the component. While robotic sensor delivery has automated the physical aspect of sensor placement, the analysis of the resulting data remains a manual and labour-intensive task. This manual analysis is not only time-consuming but also susceptible to human error, thereby limiting manufacturing capabilities. Recent advancements in machine learning (ML) and deep learning (DL) offer potential solutions to automate this analysis process.

This thesis investigates the application of DL methods to analyse Ultrasonic Testing (UT) data from CFRP composites, one of the most common NDE techniques used for aerospace components. The initial phase of research addressed the challenge of data scarcity for training DL models by utilising synthetic datasets. Various methods for generating synthetic UT data were explored, with a deep generate method resulting in a trained classifier with a 24.2% improvement in defect detection accuracy when compared to the same classifier trained on simulated data. This helped to bridge the gap between simulated and experimental training data.

Building on this foundation, the research then focused on the automatic analysis of volumetric UT data. Initially targeting defect detection, the study progressed to volumetric segmentation. For defect detection, a novel architecture was developed, achieving a 22.6% increase in classification accuracy compared to established architectures. A fully supervised model was then employed to train a 3D U-Net for segmentation, which performed well in sizing and localising

defects within the training distribution. However, the model's performance declined when tested on out-of-distribution samples.

The final phase of research explored self-supervised learning to train a model for defect segmentation, reframing the problem as one of outlier prediction. This approach eliminates the need for defective examples during training and significantly enhances generalisability. The method was also tested on an industrial component and scan setup, demonstrating promising performance and applicability in real-world scenarios.

This thesis presents promising methods to automating the analysis of UT data from CFRP composites, highlighting the potential of DL methods to improve accuracy, reduce human error, and enhance manufacturing processes in the aerospace industry.

Contents

i. List of Figures

ii. List of Tables

iii. Abbreviations

1D	-	1 Dimensional
2D	-	2 Dimensional
3D	-	3 Dimensional
AI	-	Artificial Intelligence
CAM	-	Class Activation Map
CFRP	-	Carbon Fiber Reinforced Polymer
CNN	-	Convolutional Neural Network
CV	-	Computer Vision
DL	-	Deep Learning
FBH	-	Flat-Bottom Holes
FEA	-	Finite Element Analysis
FMC	-	Full Matrix Capture
FOI	-	Foreign Object Inclusion
GAN	-	Generative Adversarial Network
GELU	-	Gaussian Error Linear Units
HPO	-	Hyperparameter Optimisation
MAE	-	Mean Absolute Error
ML	-	Machine Learning
MPL	-	Multi-Layer Perceptron
NAS	-	Neural Architecture search
NDE/T	-	Non-Destructive Evaluation/Testing
NLL	-	Negative Log-Likelihood
PNN	-	Probabilistic Neural Network
PTFE	-	Polytetrafluoroethylene
RE	-	Regularised Evolution
ReLU	-	Rectified Linear Unit
RNN	-	Recurrent Neural Networks
SHM	-	Structural Health Monitoring
SOTA	-	State of the Art
SSL	-	Self-Supervised Learning

TCG	-	Time-Compensate Gain
TFM	-	Total Focusing Method
ToF	-	Time of Flight
UT	-	Ultrasonic Testing
ViT	-	Vision Transformer
wt%	-	Percentage by Weight

1

1 Introduction:

1.1 Industrial Motivation and Research Context

Composite materials find extensive application within the aerospace, marine and civil engineering sectors, with Carbon Fibre Reinforced Polymers (CFRP) being one of the most prominent. The global demand for CFRP is expected to reach 285 kt in 2025, rising from 181 kt in 2021 [1]. Figure 1 demonstrates the strong increasing trend of CFRP use in commercial aircraft, with improved performance and efficiency in operation driven by environmental pressures. Due to their high specific strength, stiffness and corrosion resistance, composites are widely used for critical aerospace components such as wings and fuselages [2], [3]. As the use of CFRP grows, the need for effective testing of these safety critical components also increases. Ultrasonic Non-Destructive Evaluation (NDE) is the most applied method for the inspection of aerospace composites during manufacturing [4], [5]. Ultrasonic inspection is often a manual task which can be time-intensive, challenging to scale, and exposed to human factors (such as cognitive, physical or experience) which can lead to error [6].

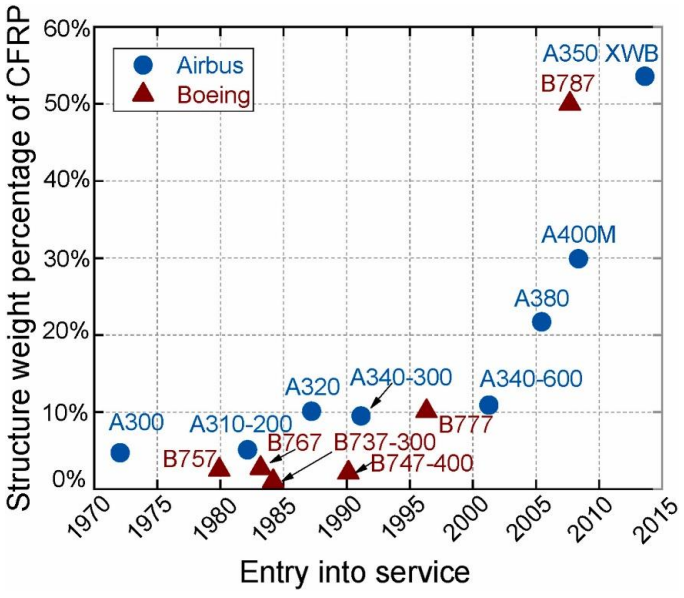


Figure 1: Increasing trend of CFRP use in commercial aircraft. Taken from [1].

The integration of robotics into NDE has enabled the efficient automation of sensor deployment for large-scale inspection processes [7]. However, despite the increased flexibility of robotic scanning and the drastic reduction in scan time seen by mechanised scanning compared to manual scanning (high degree of freedom robotics are three times faster than gantry systems [8]), the interpretation of the results in industry remains a challenging and time intensive task that requires highly trained and qualified operators to interpret results according to existing standards [9], [10], [11], [12], [13], [14]. Despite the significant improvements brought about by robotic NDE, the need for expert human interpretation of results persists. This highlights the need for further research and development of automated data interpretation techniques that can supplement or even replace human interpretation, to improve the efficiency and reliability of NDE in various industries. By reducing the dependence on human interpretation, automation can enhance the consistency, repeatability, and traceability of the NDE processes, while reducing inspection time and costs.

The interpretation of UT scan results by human operators presents two significant drawbacks, namely, poor time efficiency and the risk of human error [12]. Low levels of automation for data interpretation are feasible for mass-produced parts with precisely known geometries, but this approach typically relies on hard-coded features such as predefined time-gating, filtering, and amplitude thresholding, which may not be adequate for complex tasks, with changes in manufacturing conditions, variations in geometry, or defect characteristics [14]. DL has been identified as a key requirement for transitioning from low to high levels of industrial automation [14]. Therefore, if a Deep Learning (DL) approach could be created to automate the interpretation of complex results and work alongside the robotic inspection, the required inspection

time and quality of large components could be improved significantly, allowing for shorter signal interpretation time and a faster uptake of UT automation in aerospace and other industries. Furthermore, increased automation could lead to better defect detection capability, whilst improving consistency, traceability, and repeatability as discussed by Cantero et al. [14].

DL approaches have been identified as one of the most important enabling factors for the transition from low to high automation levels [14]. Despite the clear potential benefits of applying DL techniques to ultrasonic signal analysis for composite components, its uptake has been limited [14]. Shortage of training data is one of the main challenges that hinders research developments in this area. This shortage, combined with industrial concerns about the interpretability and compliance with standards of DL models, has presented challenges for the effective use of DL techniques. As a result, the adoption of DL in UT signal analysis for composite components has been slow, despite its promising potential to enhance the accuracy and efficiency of defect detection and characterisation.

1.2 Aims and Objectives

The objective of this thesis is to investigate the application of DL in the ultrasonic inspection of carbon fibre composites, with the aim of enhancing the automation levels observed in industrial aerospace ultrasonic testing. Considering typical tasks required by NDE operators, it is conceivable to envision an automated data interpretation pipeline, as depicted in Figure 2. The work presented in this thesis aims to breakdown and tackle different stages of the proposed pipeline.

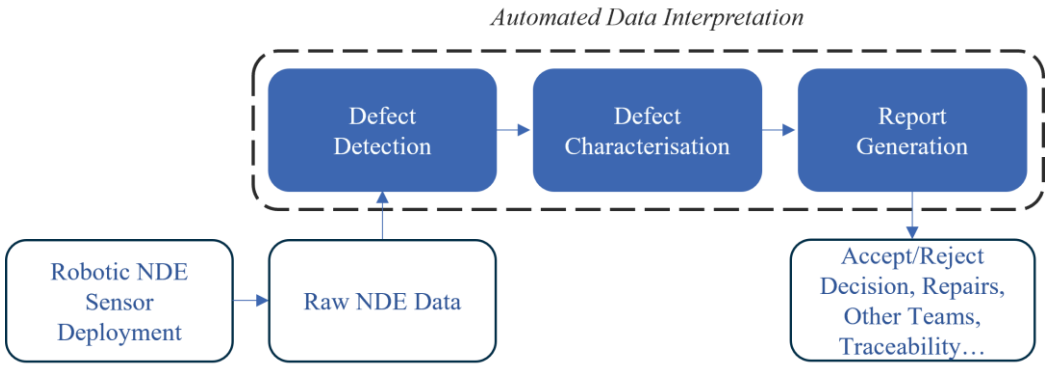


Figure 2: Suggested automated data interpretation pipeline.

1.3 Outline of Thesis Structure

The following chapters of the thesis are structured as follows:

- Chapter 2: Provides an overview of ultrasonic testing, composites, AI, along with the experimental data collection, signal processing and samples used in the following chapters.
- Chapter 3: Presents a solution to a lack of training data in NDE through the exploration of different synthetic data generation methods.
- Chapter 4: Presents work on volumetric defect detection, which utilises and expands on the work from the previous chapter to work with volumetric datasets.

- Chapter 5: Explores a supervised deep learning approach for defect segmentation in volumetric data, trying to tackle defect characterisation and report generation.
- Chapter 6: Demonstrates a self-supervised method for volumetric defect segmentation, including a case study showing industrial application.
- Chapter 7: Conclusion of the main findings of the thesis and discusses potential future work.

1.4 Contributions to knowledge

This body of work investigates DL solutions for the automated analysis of UT data for CFRP aerospace components. This thesis presents several unique and novel contributions to the field of ML in NDE. They are summarised below.

- Providing a solution to the lack of training data for NDE ML models by bridging the simulation to experimental domain gap with the development of four different synthetic data generation techniques. The evaluation and comparison of simulated and synthetic training data on experimental classification performance. Additionally, modifications to CycleGAN, including the introduction of an additional loss term and weighting the synthetic data generator more than the simulated generator, were implemented to improve the quality of synthetic data generation.
- The introduction of volumetric analysis for defect detection from UT data using DL. A new architecture was discovered which greatly improved on existing models. This was evaluated against an established VoxNet and hand-crafted architecture. Domain specific augmentation methods were also

developed and evaluated to increase model performance.

- A 3D U-Net variant was employed for volumetric segmentation to achieve accurate defect sizing and localization. The model was trained using supervised learning, with synthetic training data and defect simulation parameters providing the ground truth segmentation masks necessary for training.
- The development of a self-supervised model for volumetric segmentation of UT data. The model learns from defect-free data series along the scan direction, allowing it to establish a baseline representation. This approach enables segmentation through anomaly detection, eliminating the need for labelled defective training data.

1.4.1 Lead Author Journal Publications

1. **S. McKnight**, G. Pierce, E. Mohseni, C. MacKinnon, C. N. MacLeod, T. O'Hare, and C. Loukas, *'A Comparison of Methods for Generating Synthetic Training Data for Domain Adaption of Deep Learning Models in Ultrasonic Non-Destructive Evaluation'*. NDT & E International vol 141, no. 102978, doi: 10.1016/j.ndteint.2023.102978.
2. **S. McKnight**, C. MacKinnon, G. Pierce, E. Mohseni, V. Tunukovic, C. N. MacLeod, R. Vithanage, and T. O'Hare. *'3-Dimensional Residual Neural Architecture Search for Ultrasonic Defect Detection'*, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, doi: /10.1109/TUFFC.2024.3353408.
3. **S. McKnight**, V. Tunukovic, G. Pierce, E. Mohseni, R. Pyle, C. N. MacLeod, and T. O'Hare. *'Advancing Carbon Fiber Composite Inspection: Deep Learning-Enabled Defect Localization and Sizing via 3-Dimensional U-Net*

Segmentation of Ultrasonic Data'. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, doi: 10.1109/TUFFC.2024.3408314.

4. **S. McKnight**, V. Tunukovic, A. Hifi, G. Pierce, E. Mohseni, C. N. MacLeod and T. O'Hare, *'3-DUSSS: 3-Dimensional Ultrasonic Self Supervised Segmentation'* [Under Review].

For full research output, please see: Google Scholar Profile.

2 Literature and Background:

NDE constitutes a diverse array of methodologies employed for the inspection of components without inducing damage. Prominent techniques utilised in this domain are Radiography, Thermography, Electromagnetic methods, and Ultrasound. These methods provide inspection capabilities suitable for components of varying complexities and sizes. Each of these methods exhibits distinctive strengths and weaknesses, necessitating a detailed decision-making process when selecting a suitable approach, which often involves compromise to find the most applicable method. Factors influencing this decision encompass a multitude of considerations, including material properties, component geometry, safety considerations, resolution, implementation feasibility, operational constraints, and defect typology, among others. The intricate nature of this decision-making process underscores the complexity inherent in NDE method selection. The application of appropriate NDE techniques can significantly enhance the reliability and safety of structures and components across diverse industrial sectors.

The work conducted in this thesis focuses on the use of Ultrasonic testing. For an in-depth exploration of the various NDE methods pertinent to composite materials, readers are directed to the comprehensive investigation conducted by S. Gholizadeth [15]. This work provides valuable insights into the application of a diverse range of NDE techniques tailored specifically to the assessment of composite materials, providing invaluable insights into their respective merits and limitations.

2.1.1 Ultrasonic Testing

UT employs high-frequency acoustic waves, typically exceeding 20 kHz, to assess the integrity of components. This technique offers versatility, allowing inspection of materials ranging from metals to composites, and relies on the transmission, propagation, and reception of ultrasonic waves. UT stands out as one of the most prevalent NDE method owing to its capability for volumetric inspection, ability to detect a diverse range of defects, along with its adaptability, user-friendliness, and safety profile. Various ultrasonic-based inspection methods exist, tailored to different applications such as utilising bulk waves for volumetric detection [16] or guided waves for increased inspection ranges [17]. This thesis concentrates on bulk ultrasonic waves, primarily aimed at sub-surface defect detection.

UT has gained extensive adoption and standardisation for volumetric inspection within the aerospace industry, primarily due to its comparatively straightforward and safe implementation in contrast with radiography, alongside its capacity to identify a broad spectrum of volumetric defects [9], [15], [18], [19]. In the aerospace sector, composite UT predominantly utilises bulk wave inspection, where sound waves are excited on the surface of a component, and the reflected/scattered wave from internal scatterers can provide valuable information about the volumetric discontinuities or properties of the component.

2.1.1.1 Bulk Wave Propagation in Isotropic Medium

As waves propagate through real materials they lose energy, this effect is known as attenuation. Attenuation of a wave is due to various mechanisms such as absorption or scattering, but generally waves with longer wavelengths have lower attenuation than

waves with short wavelengths, allowing them to propagate further into a component. However, higher frequency waves with shorter wavelengths interact with smaller features allowing for higher resolution inspection. There is therefore a balance between penetration depth and inspection resolution. This balance is constrained by the relationship between operating frequency and wavelengths, as given in equation (1), where f is the operating frequency and λ is the wavelength:

$$f = \frac{v}{\lambda} \tag{1}$$

2.1.1.2 Reflection Refraction and Mode Conversion

When an ultrasonic wave meets the boundary of two media with different acoustic impedances it will undergo changes such as reflection, refraction, or mode conversion. When a wave encounters a boundary perpendicular to the interface surface, typically a portion of the wave is transmitted into the new medium, while another portion is reflected (Figure 3). This is given by T and R respectively in equations (*2*) and (*3*). Here, Z denotes the acoustic impedance of a given material. Acoustic impedance refers to a materials resistance to the propagation of ultrasound. This is given by equation (*4*), where ρ is the material density. Consequently, when interacting with materials exhibiting a substantial disparity in acoustic impedance, the wave will undergo a greater degree of reflection.

$$R = \frac{Z_2 - Z_1}{Z_1 + Z_2} \tag{2}$$

$$T = \frac{2Z_2}{Z_1 + Z_2} \tag{3}$$

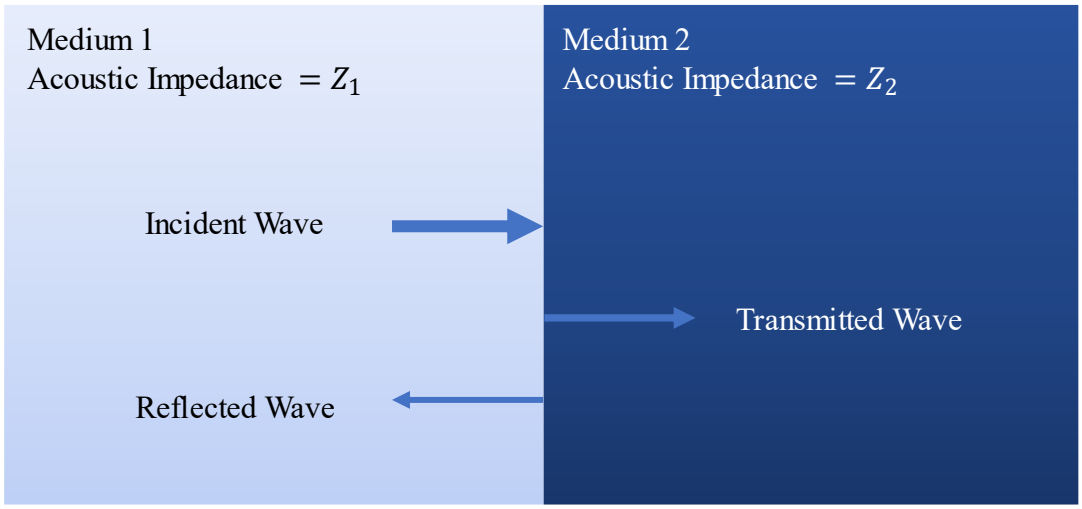


Figure 3: Illustration of reflected and transmitted waves for the interaction of an incident wave normal to the boundary of different mediums.

$$Z = \rho v \tag{4}$$

When the acoustic wave does not meet the interface boundary at a normal angle, a change of direction is observed (Figure 4). This is known as refraction and is defined by Snells law, given in equation (5).

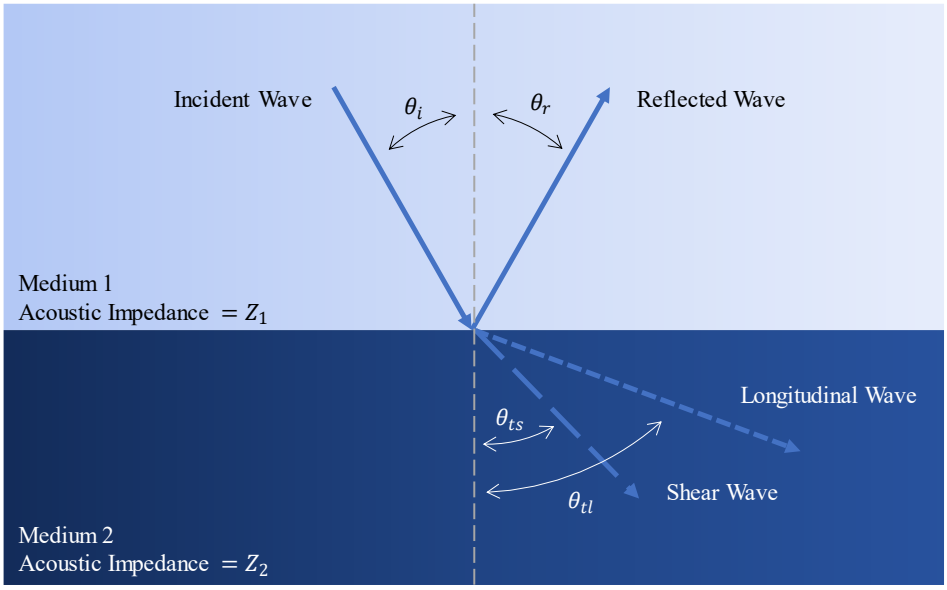


Figure 4: Illustration of reflected and transmitted waves for the interaction of an incident wave at an angle to the boundary of different mediums.

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_{rl}}{v_{rl}} = \frac{\sin\theta_{ts}}{v_{ts}} = \frac{\sin\theta_{tl}}{v_{tl}} \quad (5)$$

When the refracted angle $\theta_{tl} < 90°$ (equation (5), Figure 4), known as the critical angle, mode conversion occurs. Mode conversion is the propagation of a different wave type e.g. longitudinal waves create shear waves. At the critical angle ($\theta_{tl} = 90°$), the incident longitudinal wave is converted to a surface following longitudinal wave. Beyond the first critical angle ($\theta_{tl} > 90°$) the longitudinal wave is totally internally reflected, meaning it does not pass into the second medium but is instead reflected back into the first medium, and only the shear wave is refracted into the material.

While numerous other wave types and factors influence ultrasound, this section offers an overview of the primary macro components contributing to bulk waves, which constitute the majority of ultrasonic inspection [20].

2.1.1.3 Single Element Ultrasound

A single piezoelectric ultrasonic transducer is the simplest means to convert electrical current to ultrasonic waves and vice versa. A single transducer can be used to inspect the internal volume of a component along with other tests such as thickness measurements using pulse-echo. In pulse-echo, the transducer is attached to the surface of a component, often with a coupling medium (or couplant), which aids in removing the air surrounding the components' surface, facilitating acoustic energy transfer from the transducer to the test component. An acoustic wave is then propagated through the surface of the component, traversing the thickness and reflecting off the backwall surface in a healthy component (this can be used for thickness measurements if the

speed of sound in the component is known) or off a discontinuity in a defective component (Figure 5). The reflection at the surface of the component and any later reflections are recorded using the transducer which converts the pressure into a voltage. This is often visualised in the form of an amplitude versus time (A-scan) plot. Pulse-echo inspection sees wide applicability to industrial settings where access is limited to a single side of a component.

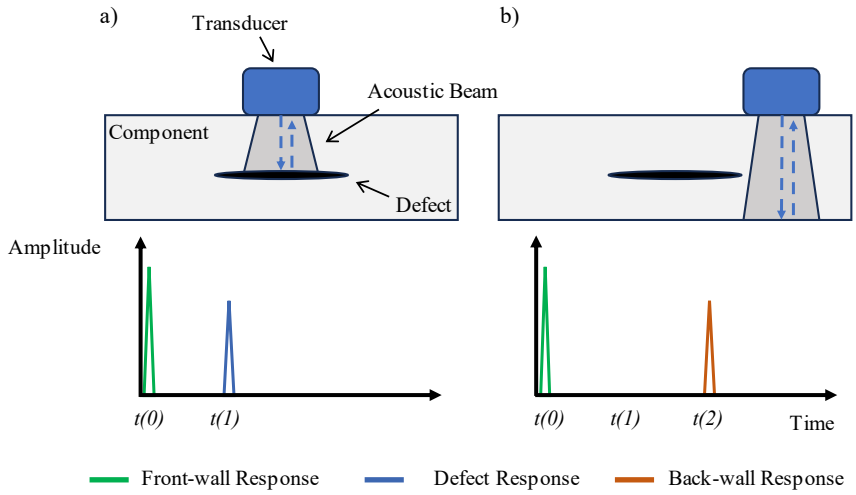


Figure 5: Example of pulse-echo inspection for a defect response (a) and defect free response (b).

Through-transmission is a common alternative to pulse-echo which makes use of two transducers, one for transmission and one for reception, on either side of the component (Figure 6). However, this method requires access to both sides of the component and requires good alignment of the two probes. This can make it more challenging to implement in industrial settings than the pule-echo method.

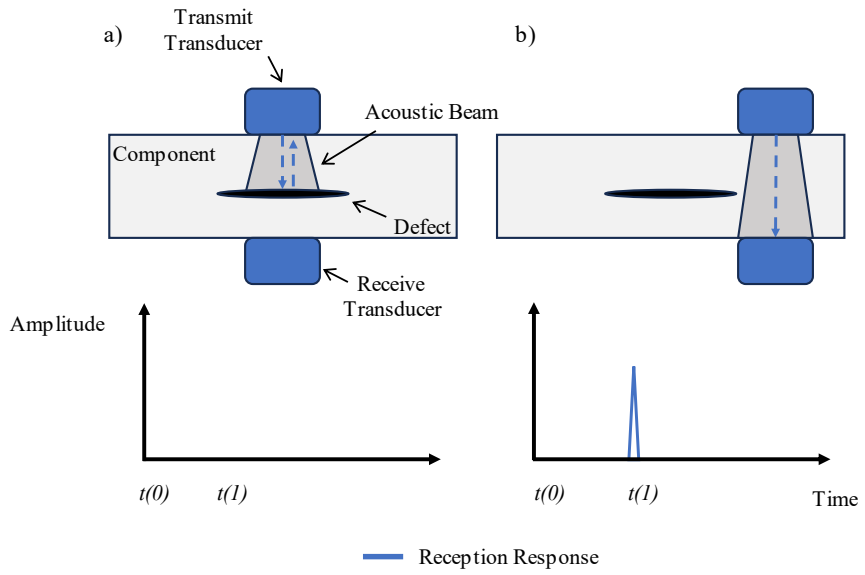


Figure 6: Example of through-transmission inspection for a defect response (a) and defect free response (b).

2.1.1.4 Phased Arrays, Mechanised Scanning and Data Representation

The use of Phased Array Ultrasonic Testing (PAUT) has become increasingly popular to generate and receive ultrasonic sound waves owing to their operational flexibility. Phased arrays employ independently controllable UT elements that enable the collection of richer datasets through more complex electronic scanning and imaging methods, such as beam steering, dynamic depth focusing, and variable sub-apertures [21]. Examples of how firing delays can generate different beams can be seen in Figure 7. Recently Full Matrix Capture (FMC) has become increasingly popular as it collects the full set of transmit/receive element combinations which allows for different imaging methods to be applied post-acquisition such as the Total Focusing Method (TFM) [22]. When constructing a TFM image a delay and sum approach is used, where the known speed of sound in the material is used to give the expected time index of each FMC transmission/reception pair. For every pixel in the image the sum of each

pairs corresponding time indexed amplitude gives the total magnitude of the response. This leads to an increased signal to noise ratio and improved focusing within the image. Such imaging methods work very well for isotropic materials with consistent and well-defined acoustic velocities. However, for anisotropic medium with variable speeds of sound, the calculation of appropriate time of flights becomes more challenging and can be intractable for certain scenarios without the use of simulations [23], making these advanced imaging methods often inappropriate for industrial inspections, where efficiency is a key driver.

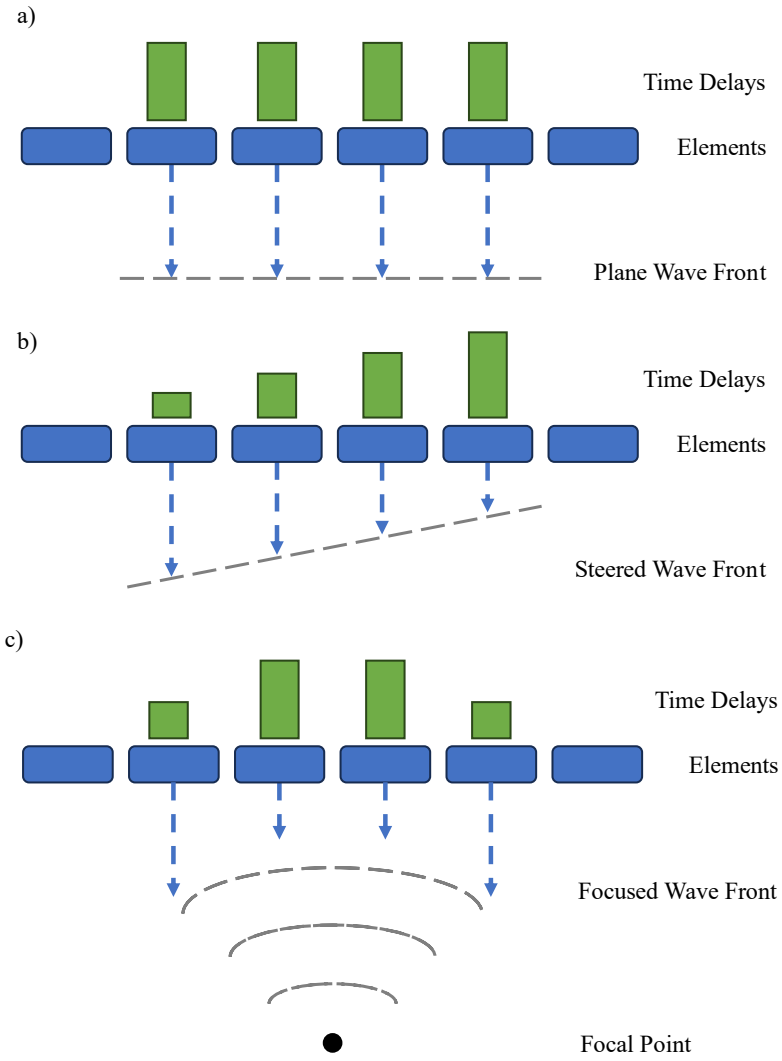


Figure 7: Illustrations of standard array scanning methods: (a) Plane wave inspection, (b) steered inspection, (c) focused inspection.

By controlling each individual element (or sub-aperture of elements) of a linear phased array, depth-wise sectional images (B-scans) can be created in a single scan (Figure 8, Figure 9). When combined with mechanised scanning perpendicular to the length of a linear phased array, complete 3-dimensional (3D) volumetric scan data of components can be generated by stacking multiple individual B-scans together at known positions (Figure 9 (b)). Mechanised scanning was first introduced with gantry or bridge-based systems which aimed to automate sensor delivery for simple or fixed geometries[24].

The adoption of more complex high axis industrial robots with greater degrees of freedom has led to increased flexibility [8], [24], [25], [26], with previous systems demonstrating impressive scanning rates of 25.3 square meters per hour [8]. The integration of robotics and PAUT into NDE has revolutionised large-scale inspection processes by enabling efficient automated inspection of large, complex components [7]. It allows for increased flexibility, repeatability and the drastic reduction in scan time seen compared to previous manual scanning approaches and has significant positive implications for the reliability and safety of aerospace structures.

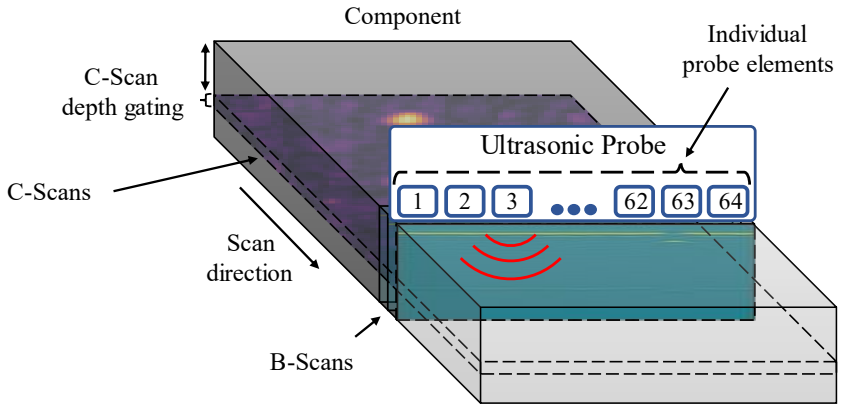


Figure 8: Demonstration of how individual elements construct a linear phased array to produce B-scan and C-scan images when inspecting a component.

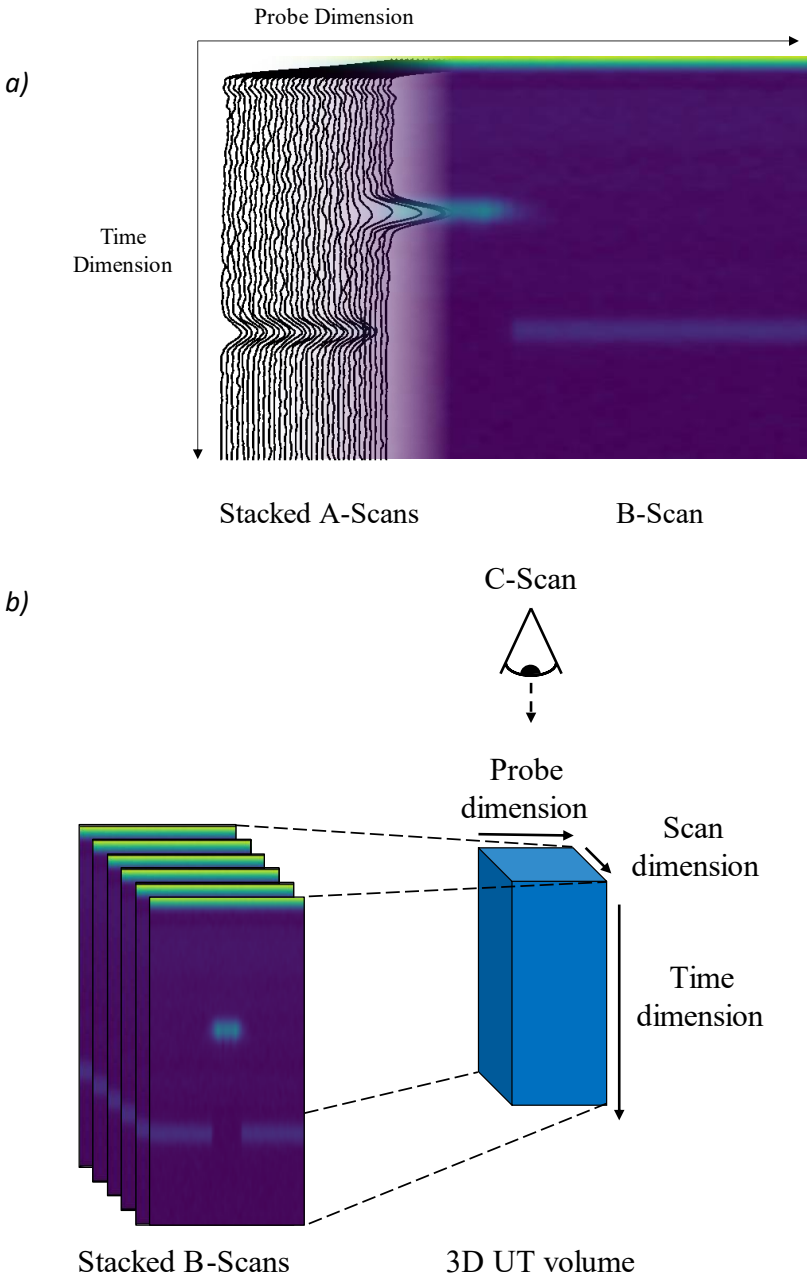


Figure 9: a) Representation of how A-scans are stacked to form B-scans. b) How B-scans are stacked to create a full UT volume.

UT data is commonly visualised as images, either by selecting a B-scan directly, or as an amplitude or time of flight C-scan; where either the maximum response amplitude (examples can be seen in Figure 22) or the time index of the maximum response amplitude within the volume is imaged to produce a top-down section view across the

sample. B-scan imaging offers the most comprehensive understanding of the volumetric response of a sample, yet it sacrifices comparative spatial information along the scan axis, which can make it impractical for defect sizing and inefficient for the complete analysis of large components. Conversely, C-scans maintain comparative spatial information but at the expense of compressing the temporal dimension and losing depth-wise information. As highlighted by Smith et al. [27], who extracted additional information such as ply orientation for composite inspection from C-scans, which is not readily available from B-scans. Typically, practitioners employ both imaging methods concurrently to evaluate components effectively and efficiently, as certain imaging methods are appropriate for different defects, as given in Table 1.

Table 1: Appropriate UT imaging method for a selection of different composite defects, with applicability levels ranked in ascending order as: none, low, medium, high, v. (very) high. Modified from [28].

Defect	A-Scan	B-Scan	Amplitude C-Scan	Time-of-Flight C-Scan
Delamination (<10 mm)	Medium	High	High	High
Delamination (>10 mm)	High	V. High	V. High	V. High
Disbond	High	V. High	V. High	V. High
Void	Medium	Medium	V. High	Medium
Impact	V. High	V. High	V. High	V. High
Porosity	Medium	Medium	High	Medium
Inclusion	High	High	High	Medium
Fiber Waviness	Low	High	High	None
Fiber misalignment	None	Low	High	None

2.1.2 Composites in Aerospace

Composites are versatile, often anisotropic materials that are widely used in many industries due to their favourable properties such as corrosion resistance, high specific strength, and high specific stiffness. Layered composite structures are generally anisotropic. This anisotropy allows for precise engineering to meet specific structural

requirements, making these composites ideal for high-performance applications [9], [18], [19], [21], [29], [30], [31], [32], [33]. CFRP is a widely used layered composite in the aerospace industry making up over 50 percent by weight (wt%) for the two most recent long-range aircraft, the Airbus A350 and the Boeing 787, and up to 70-80 wt% for private jets and helicopters [1]. CFRP is manufactured by layering multiple carbon ply sheets, which are subsequently cured following the addition of a thermoset polymer.

It is necessary for enhanced automation in composite manufacturing to improve efficiency and mitigate costs. However, the manufacturing process of composite components is complex which can introduce defects, compromising their integrity and performance [15], [18], [21], [29], [31], [33], [34]. Given that rejected components result in heightened expenses due to rework and repairs, automated testing also emerges as a pivotal requirement. Defects can range from delamination and cracks to foreign object inclusions, fibre distortions, and porosity [19], [34] (for further details please refer to section 2.1.2.1). These defects represent an increased risk as they are often not detectable by the naked eye and can affect strength and fatigue behaviour [33]. Cyclic stresses from operation can cause these defects to grow to a critical level where they may cause catastrophic failure of the component [31]. As the use of composites in safety-critical parts continues to rise, the detection, characterisation, and quantification of defects become increasingly important [31]. Moreover, as the automated production of aerospace components continues to expand, so does the requirement for automated testing. [33]. Automated NDE techniques are therefore well placed to inspect these components as they can provide information about the integrity of the structure below the surface at the scales required.

Intensive NDE is widely performed on aerospace components during manufacturing and assembly [19]. However, some of the physical properties that make composites so advantageous for use in components, also add significant challenges for NDE inspection [19], [29], [30]. The anisotropic and inhomogeneous nature of composites makes NDE far more challenging than compared to more traditional materials, such as metals [21], [29]. For example, in ultrasonic NDE, wave propagation in anisotropic composite structures is complex, with variability in scattering due to "macro" structural features such as material lay-up and "micro" non-structural features such as local anisotropy. This can lead to high attenuation of ultrasonic waves, reducing the signal to noise ratio and in turn reducing the probability of detection [21], [32], [35]. In certain cases, this can pose challenges in distinguishing whether irregularities detected during testing represent defects or variations in the base material [34]. Furthermore, accurately determining the ultrasound velocity within the sample adds another layer of complexity. This parameter varies depending on specific factors such as the resin type, fabric stacking sequences, and fibre orientation. Additionally, the customisable anisotropic nature of composites, which can be tailored to meet design specifications, further complicates matters. These intricacies introduce varying levels of inconsistency across different components, posing significant complexities for the NDE of composite materials.

2.1.2.1 Defects

The work conducted in this thesis focuses on defects occurring during manufacturing, where a wide range of defects can be introduced [19]. However, damage can also be caused during the service-life of a component from impacts, loading cycles etc. This damage can often lead to cracking in cured samples which results in potential failure

modes that are challenging to predict due to the materials' anisotropic characteristics. Further details on the specific defect types commonly seen in manufacturing are outlined in this section.

2.1.2.2 Voids and Porosities

Voids and porosities are significant and common defects commonly encountered in composite materials, which present significant challenges to their structural integrity and performance [36]. Voids refer to isolated trapped air pockets or gas bubbles within the material matrix, while porosities are microscopic pores which are often dispersed throughout the composite structure. These defects can arise during the manufacturing process due to incomplete resin impregnation, improper curing conditions, or inadequate vacuum or pressure application during fabrication. Additionally, voids and porosities can also result from the presence of contaminants or moisture in the composite materials. Voids can often be detected ultrasonically since the significant mismatch in acoustic properties between air and the composite material results in a significant amplitude response. The size of porosities makes them far more challenging to detect than voids, it is also very challenging to eliminate them completely from manufacturing and most parts will have an allowable porosity percentage, for primary aerospace structures this is often less than 2% [37]. Porosities therefore rarely give a strong amplitude response ultrasonically and have to be detected due to an increased attenuation, often determined as a lack or reduction in back wall response. The presence of voids and porosities can compromise the mechanical properties of composites, including strength, stiffness, and fatigue resistance, as they create stress concentration points and reduce the effective load-bearing capacity of the material.

2.1.2.3 Delaminations

Delaminations represent a critical concern in composite materials, characterised by the separation or splitting of layers within the laminate structure. They are one of the more common defects found in composites [38]. These interfacial defects can arise during the manufacturing process due to insufficient bonding between layers, voids, or resin-rich regions. Delaminations can also occur because of mechanical loading, impact damage, or environmental factors such as moisture absorption. Delaminations can range in size and as a result have a ranging impact on a component from negligible to severe. Left undetected or untreated, delaminations can propagate, leading to significant reductions in structural integrity and mechanical performance. As delamination's run parallel to the ply orientation and subsequently the surface of the material, they are often well detected by ultrasonic testing due to their significant amplitude response. However, in certain exceptional cases, plies may maintain contact without the ability to transfer strain, a phenomenon commonly known as a "kissing bond" [39]. Owing to the absence of separation between layers, detecting these specific delamination's can prove particularly challenging.

2.1.2.4 Foreign Object Inclusions

Foreign Object Inclusions (FOI) encompass any unintended foreign materials embedded within composite structures during manufacturing or service, and can include particles, fibres, or debris from processing equipment [40]. These inclusions may lead to localised stress concentrations, delamination, or initiation of cracks, which diminish the overall strength and durability of the composite material. The detection of FOIs within composite materials is influenced by the diverse properties of the

inclusion materials. For instance, fibrous inclusions may pose challenges for ultrasonic detection because their acoustic properties closely resemble those of the surrounding medium after curing. In contrast, FOIs with markedly distinct acoustic properties from the composite matrix can be readily identified. This variability in detectability underscores the importance of understanding the acoustic characteristics of different inclusion materials and employing appropriate inspection techniques tailored to their specific properties. In practice, the manufacturing of high-value components often takes place within controlled environments such as clean rooms to minimise the occurrence of FOIs. These sterile settings aim to reduce the introduction of contaminants and foreign materials during the manufacturing process. However, despite stringent measures, it is impossible to completely eliminate the risk of FOIs.

2.1.2.5 Fibre distortions (or marcels)

Fibre distortions can occur during the manufacturing process, where fibres may experience misalignment, waviness, or kinking, compromising the intended mechanical properties of the composite [41], [42]. Fiber distortions can result from various factors, including improper handling, resin flow issues, or inadequate consolidation during curing. These distortions can significantly impact the strength, stiffness, and fatigue resistance of the composite, as they introduce weak points and stress concentrations along the fibre-matrix interface. Additionally, fibre distortions can affect the uniform distribution of load-bearing capabilities within the composite structure, leading to non-uniform mechanical performance.

2.1.2.6 Resin Rich/Starved Areas

Resin-rich regions occur when excess resin accumulates within the composite, leading to non-uniformity and potential weakening of the material. Conversely, resin-starved areas occur when there is insufficient resin to fully saturate the reinforcing fibres, resulting in inadequate bonding and reduced load-bearing capacity [43]. These issues can arise during the manufacturing process due to improper resin infusion, resin flow, or vacuum pressure. Resin-rich areas may lead to increased weight, reduced stiffness, and potential delamination, while resin-starved regions can result in reduced strength, increased susceptibility to cracking, and compromised durability. Quality control measures, such as optimised resin infusion techniques and careful monitoring of resin-to-fibre ratios, can help mitigate these issues during manufacturing. Changes in ultrasonic attenuation can be used to evaluate resin rich or starved areas, as these impact the components fibre-volume-fraction which has a direct impact on acoustic impedance.

2.1.3 Ultrasonic Testing of Composites

UT has been widely adopted and standardised for testing in the aerospace industry due to its ease of implementation and ability to detect a wide variety of defects [9], [15], [18], [19]. The typical use for UT is with normal incident longitudinal waves so that wave propagation is independent of ply orientation [19]. The interaction of longitudinal waves and ply thickness is normally weak due to the wavelengths at typical testing range of single megahertz frequencies being much greater than the typical ply thicknesses [19]. The National Composites Network released a best practice guide for applying NDE to composites [44]. This includes what inspection methods are best suited for different types of defects. At the time of writing, UT pulse-echo had

been proven to detect the greatest number of flaws. Hsu et al. and Bossi and Georgeson [19], [45] discussed specifically what defects UT can detect in aerospace composites, and the differences in ultrasonic flaw interactions of each defect type. They also suggested the most appropriate methods for detection of different defects. When inspecting composite UT data inspectors will often use both time of flight and amplitude C-scans simultaneously to aid with inspection [45].

Whilst many research papers rely on typical manufactured defects, Kokorov et al. [32], aimed to simulate typical manufacturing defects more accurately in composites, where the structural defects have similar physical and chemical properties as the binding material. Their results demonstrated that UT can efficiently detect defects which have similar physical and chemical properties to the composite binding material, helping to bridge the gap between the detection capability of naturally occurring and manufactured defects.

2.1.4 Artificial Intelligence

2.1.4.1 Fundamentals

Artificial Intelligence (AI) serves as an overarching concept encompassing machines capable of leveraging knowledge and addressing diverse problem-solving tasks that typically require human intelligence [46], [47]. There has been significant historical research into different approaches for achieving AI. Early efforts in AI, particularly between the 1950s and 1990s, focused heavily on symbolic reasoning and rule-based systems. These paradigms formed the foundation of symbolic AI, a branch of AI where knowledge was explicitly encoded into structured rules and logical relationships. This approach aimed to model intelligence through explicit representation of knowledge,

enabling reasoning and decision-making in domains with clearly defined parameters [48]. These systems demonstrated significant promise in early AI research for well-defined tasks, with notable examples such as IBM's Deep Blue [49], a chess-playing program, and ELIZA [50], the first chatbot. However, while rule-based approaches excelled in structured environments, they often fail when encountering situations which do not match their heuristics, which can often be the case for the complexity and variability of real-world scenarios [51]. Model-based approaches attempt to address these problems by basing solutions upon a theoretical model of a component or system [52]. They are less dependent on expert opinion and offer higher flexibility and scope for expansion. However, these systems can become expensive and complex to construct. Symbolic AI has faced varying challenges such as the manual effort required for knowledge acquisition, the rigidity of rule-based frameworks, challenges with scalability, and its inability to effectively handle perception tasks like image and speech recognition.

To overcome these challenges, the late 1980s and early 1990s marked a pivotal shift toward connectionist approaches [53], which emphasized data-driven learning through artificial neural networks. Unlike symbolic systems, which relied on predefined rules, connectionist methods allowed machines to learn patterns and representations directly from data, making them more adaptable and robust in handling noisy or incomplete information. This shift was aided by advances in algorithms, particularly the development of backpropagation [54], which facilitated the training of multi-layer networks. Connectionist models proved especially adept at tasks requiring perception and pattern recognition, areas where symbolic AI struggled. Concurrently, advances in computing power and the growth of the global internet provided the computational

resources and vast datasets needed to train more complex and capable models, further accelerating the adoption of connectionist paradigms.

While connectionist approaches gained significant traction, they did not entirely replace other AI methodologies. Instead, the field expanded to incorporate a range of techniques that remain relevant today. These include fuzzy logic for reasoning under uncertainty, rule-based systems for applications requiring interpretability, model-based reasoning for leveraging structured system models, evolutionary algorithms inspired by natural selection, hybrid systems, and data-driven machine learning (ML) methods [55], [56], [57].

ML constitutes the broad subset within the field of AI, emphasising the development of algorithms and models capable of learning from data (such as connectionist approaches) (Figure 10). This often allows for solving complex problems where rules-based approaches would be intractable. ML problems can be categorised into either classification or regression tasks. In classification tasks, the objective is to categorise data into distinct classes or groups, whereas in regression tasks, the aim is to predict continuous numerical values. DL is a subset of ML (Figure 10) as brought about by the connectionist movement which, loosely inspired by the human brain, makes use of neural networks with at least one hidden layer. In general, DL is employed to tackle more intricate and demanding tasks compared to alternative ML methods. However, achieving success in DL often necessitates access to substantial amounts of data. DL is widely used and has been successfully applied to many challenging tasks such as health care (diagnostic assistance, drug discovery, virtual healthcare etc. [58]), natural

language processing (translation, summarisation etc. [59]), or Computer Vision (CV) (pose estimation, depth estimation, autonomous navigation etc. [60]).

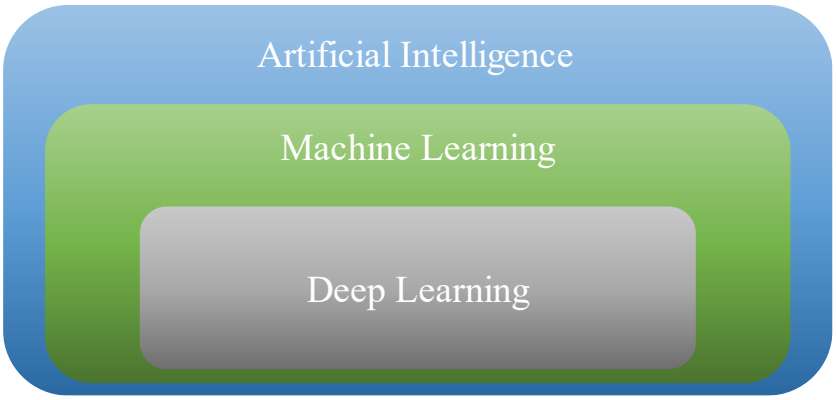


Figure 10: Illustration of the nested relationship between AI, ML and DL.

When trying to implement a ML solution there are many factors to consider. These can be broken down into distinct areas: 1) the model which provides the solution, 2) the data which allows the model to improve its performance, 3) a training regime to effectively transfer the knowledge from the data onto the model. There is not a one-size-fits-all method for every task and finding the right approach for each domain is not straightforward. Variations in one aspect often influence other areas, making the process complex and non-trivial.

The massive increase in data availability and compute resources over the last couple of decades has led to an explosion in the use of DL and has provided state-of-the-art (SOTA) solutions to tasks previously thought unsolvable, such as complex perception tasks like image processing. In the context of NDE, DL offers distinct advantages over traditional AI approaches like symbolic AI. While symbolic AI relies on predefined rules and logical reasoning, it can struggle to handle unstructured data, such as images, signals, and sensor outputs, which are common in NDE applications. Symbolic systems are also less flexible when faced with the variability and noise inherent in real-

world data. On the other hand, DL models, can learn features from raw data, enabling them to adapt to complex, noisy, and high-dimensional inputs without the need for exhaustive manual feature engineering which can be highly application specific and require extensive expert intervention. This makes DL particularly effective for defect detection and classification in materials, where patterns can be subtle and unpredictable. Given these advantages, this thesis primarily concentrates on DL solutions. Consequently, the following sections offer an overview of DL to provide necessary context and understanding.

2.1.4.2 Deep Learning

The perceptron, introduced by McCulloch and Pitts in 1943 [61] and first demonstrated by Frank Rosenblatt in 1957 [62], is the basic building block for the original neural networks and was inspired by workings of biological neurons in the brain [63]. The functioning mechanism involves taking inputs, summing them following a linear transformation using weights along with the addition of a constant term, known as a bias. Subsequently a non-linear transformation is applied which facilitates backpropagation and allows for stacking of multiple layers of neurons, this is often referred to as an activation function (Figure 11). By combining multiple perceptron's together to produce Multi-Layer Perceptron's (MLPs) it is possible to approximate any function [64]. Advances in neural network research has introduced modifications to the perceptron, but the underlying principles are the foundations for most neural networks.

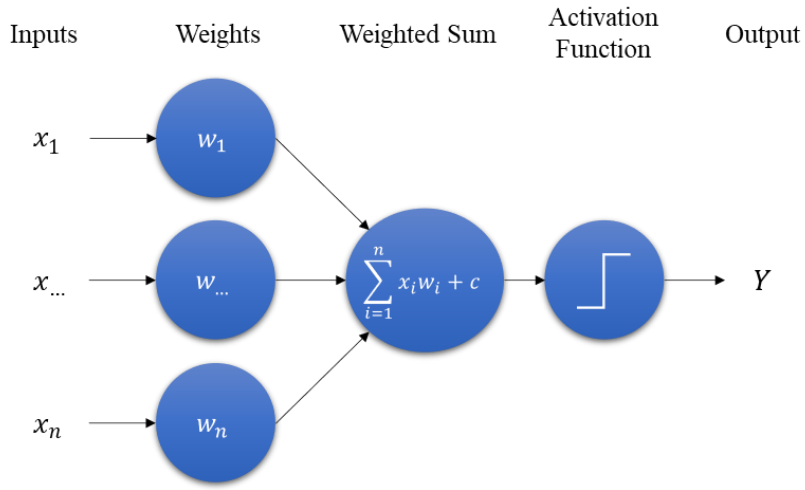


Figure 11: Illustration of a perceptron.

DL encompasses a range of different neural network architectures which all have at least one hidden layer [65]. Traditionally, these architectures faced constraints on the number of layers due to challenges in information propagation through deep networks, a phenomenon known as the vanishing gradient problem. [66]. However, as research into deeper networks continued solutions to the vanishing gradient problem emerged [67]. This facilitated the construction of significantly deeper and larger neural networks, empowering them to learn far more intricate features and undertake more complex tasks.

Many different network architectures exist, and they are often task dependent, with the construction of the best network often a challenging part in the process in applying DL. However, despite the near infinite array of different architecture combinations, there are a few key principles that underpin the training of these models. These are forward pass, network loss, and back propagation.

Forward pass is the process of passing data through a model, crucially this can then be evaluated using some metric known as loss. The network loss is the measure of how

well our model fits to the training data. A model with a high loss hasn't learnt much information from the training data whilst a model with a low loss is capable of mapping to the training data very well. There are multiple different types of loss functions used in training DL models, for details on specific loss functions please refer to the relevant literature [68], [69].

After the model's performance is assessed using the forward pass and a loss value is calculated, backpropagation is then used to adjust the model to better fit to the training data. This generally involves adjusting the model parameters using one of many optimisation algorithms based on the gradients computed during backpropagation. The field of optimisation methods for DL models is a large one with many different approaches available. The Adam optimiser [70] has proven to be highly effective in many cases [67], [71]. Nonetheless, numerous alternative options exist, and once again, the choice of optimiser is often contingent on the specific application. For additional insights on model optimisers, please consult the available literature. [72], [73].

This iterative learning process is generally repeated multiple times during training until a model is produced that has learnt some representation from the training data. Typically, a model is evaluated during training against a separate dataset known as the validation set. The validation set comprises data that is from the same distribution as the training data, but the model has not been exposed to during training. This evaluation allows for monitoring the model's performance on unseen data, providing insights into its generalisation capabilities. By assessing performance on the validation set, training can be stopped at an appropriate point, preventing overfitting to the

training data, and ensuring that the model performs well on new, unseen data. This approach helps to ensure that the model learns meaningful patterns from the data, rather than simply memorising the training examples. Once training is completed final model performance is evaluated on a separate, unseen test set. This ensures there is unbiased reporting of the final model performance.

2.1.4.3 Computer Vision and Convolutional Neural Networks

CV is the field of using computers to interpret and draw information from visual data. Over the last couple of decades DL has revolutionised the field of CV [74], [75]. There are many different CV applications, but they can generally be categorised into classification, object detection, and segmentation tasks. In classification tasks, models aim to extract features from images or video data and categorise them into distinct groups based on the visual information. For object detection the goal is to localise specific objects within an image or video, generally by drawing bounding boxes around them. This process integrates classification within an image and combines it with approximate instance localisation. For segmentation type tasks the objective is to classify each pixel within an image, thus not only achieving classification but also highly accurate localisation (Figure 12). This can be extended to instance segmentation where individual objects are distinguished from one another, enabling precise delineation and identification of each object instance within the image. Considering that ultrasonic data is often visualised as images, comparing these tasks with the automated ultrasonic inspection pipeline (Figure 2) underscores a notable alignment. This implies that computer vision methodologies could offer potential solutions to many of the challenges in automated UT data analysis.

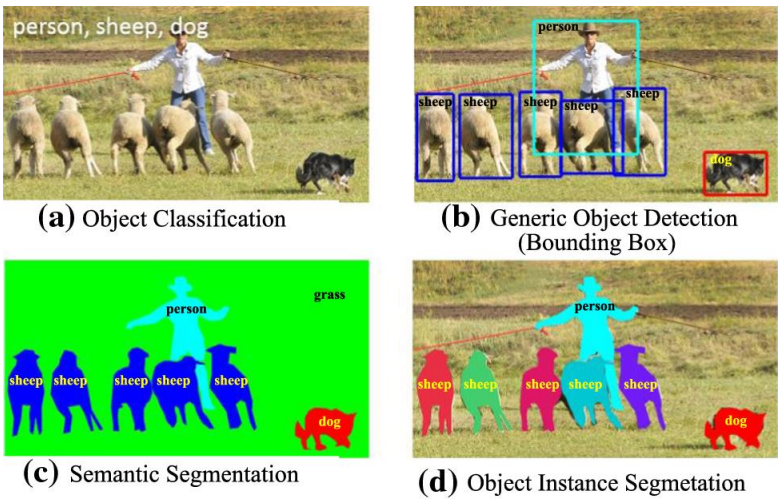


Figure 12: Classical computer vision problems. (a) Image level classification, (b) Object detection, (c) pixel wise semantic segmentation, (d) instance level semantic segmentation. [76]

As previously discussed, DL encompasses a wide array of different model architectures, but in CV applications Convolutional Neural Networks (CNNs) have repeatedly demonstrated wide scale success in image classification and other CV tasks [74].

CNNs make use of convolutional operations, which are built into convolutional layers to extract information from images (Figure 13). Convolutional layers are often combined with down sampling layers for dimensionality reduction and activation functions to construct complete network architectures (Figure 14).

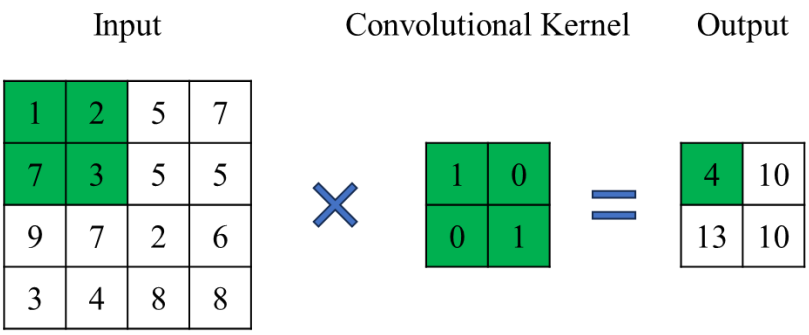


Figure 13: Illustration of a convolutional filter.

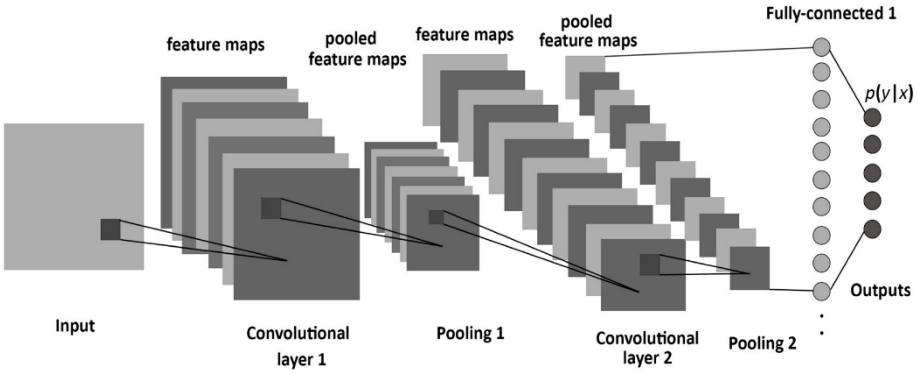


Figure 14: The structure of a generic CNN, consisting of convolutional, pooling and fully connected layers [77].

Over the past decade CNNs have evolved significantly with changes in architectures, training methodologies, data availability, and increased computational power all contributing to improved results. 2012 marked a significant milestone in the evolution of CNNs with the emergence of AlexNet [78]. AlexNet utilised deep convolutional layers, overlap pooling, Rectified Linear Unit (ReLU) activation functions, and dropout for regularisation. This was the first demonstration of the remarkable potential of CNNs at scale, significantly improving on previous results for the ImageNet Large Scale Visual Recognition Challenge [79]. This achievement underscored the efficacy of CNN methodologies in the domain of CV, triggering a surge of interest and subsequent advancements in CNN architectures and applications. After the introduction of AlexNet, the field witnessed notable progressions, exemplified by the

advent of VGG in 2014 [80]. VGG introduced novel architectural features such as smaller convolutional filters coupled with deeper network structures, thereby enhancing the capacity of CNNs to extract intricate features from visual data. VGG's simplicity in using uniform layers made it easier to understand and implement, although it came with increased computational cost due to its depth and large number of parameters. In 2015, the introduction of ResNet [67] presented a breakthrough by introducing skip connections, which helped mitigate the vanishing gradient problem. Prior to the introduction of ResNet the vanishing gradient problem was limiting the depth of CNNs being trained, which in turn limited their ability to learn complex data representations. The vanishing gradient problem refers to the phenomenon in deep neural networks where gradients calculated during backpropagation diminish exponentially as they propagate backward through multiple layers of the network. This diminishing gradient effect can hinder the training process by making it difficult to update the weights of early layers effectively, leading to slower convergence or even stagnation in learning. Skip connections enabled the gradient to flow directly through the network, making it feasible to train exceedingly deep networks comprising hundreds of layers and thus capture more complex features.

CNNs continued to advance, with competition from other architectures such as transformers providing an alternative method and often achieving near state-of-the-art results in large visual recognition tasks [81]. Vision Transformers (ViTs) represent a departure from traditional CNNs by leveraging transformer architecture, originally popularised in natural language processing. ViTs process images as sequences of patches, using the self-attention mechanisms to capture global and local dependencies within these patches. This approach has shown promising results in various computer

vision tasks, demonstrating its potential to compete with and sometimes outperform CNNs, particularly in tasks requiring long-range dependencies and global context understanding. However, ViTs generally demand larger amounts of training data compared to CNNs due to their dependence on self-attention mechanisms, which require sufficient diverse examples to effectively learn both global and local dependencies within image patches. In contrast, CNNs traditionally rely on local receptive fields and hierarchical feature extraction, which can sometimes generalise well even with smaller datasets. Therefore, while ViTs have shown promising results, particularly in tasks demanding global context understanding, their performance often hinges on the availability and quality of extensive training data.

Despite the impressive performance and popularity of ViTs, in 2022 Liu et al. introduced ConvNext [82], reaffirming CNNs' ongoing effectiveness in computer vision tasks. ConvNext demonstrated how different architectural modernisations could improve a standard ResNet to compete with Swin Transformers (a form of ViT) [83]. This work brought together and analysed the impact of several strategies for modernising CNNs (often inspired by ViTs). This included the integration of skip-connections, separate down sampling layers, the exploration of alternative and fewer activation functions such as Gaussian Error Linear Units (GELU), a reduction in normalisations and the use of layer normalisation instead of batch normalisation, and experimentation with varying kernel sizes. This work resulted in a significant advancement in the capabilities of CNNs, showcasing their adaptability and competitiveness against transformer-based architecture.

It is evident that, akin to the broader landscape of DL, the architecture of modern CNNs encompasses a diverse array of configurations and design choices, often tailored to the specific requirements of individual tasks [84], [85]. This diversity stems from fundamental differences in the design principles and components of these architectures, which significantly impact their performance and suitability for various computer vision challenges.

2.1.4.4 Data

ML systems depend on data to learn from during training. Consequently, a substantial amount of high-quality training data is crucial in the development of ML, especially DL models as they require large amounts of data to effectively update numerous parameters and uncover complex solutions to tasks. The availability of large open-source datasets such as CIFAR-10 [86] and ImageNet [79] has enabled massive growth in CV and has contributed significantly to the enhanced accuracy of models.

Zhu et al. [87] evaluated how the size of a training datasets impacts the quality of object detection models. They concluded that simple models become saturated surprisingly early and are unable to benefit from very large datasets. However, more complex models can make use of greater features and benefit from significant datasets (potentially up to 10^{12}). They go on to suggest that the relationship between model complexity and saturation of datasets may be logarithmic; suggesting that as model complexity increases datasets will need to increase exponentially before the model becomes saturated. However, the returns from super large datasets quickly diminish due to the logarithmic nature of the performance increase.

Data augmentation is a valuable technique to try and get more out of limited training datasets by modifying the original dataset. For images, there exist a wide range of augmentation methods from geometric transformations, such as flipping, rotations, translation, cropping or colour space adjustments, to image mixing or random erasing [88]. Whilst, many different augmentation methods are available, the selection of the appropriate method/methods is dependent on the specific application area, as not all methods will always be suitable. For example, random erasing is not always a label-preserving transformation and may make the image unrecognisable. In an NDE context, if you were to erase the defect signature from an image, the image would not represent the original defect classification. Therefore, this method may not be applicable or would require adjustment/manual intervention such as object-aware random erasing.

2.1.4.5 Summary

AI, ML, and DL are broad and complex fields with a wealth of research and applications that cannot be fully covered in this thesis. This section introduces key concepts and terms relevant to the scope of this work. AI aims to enable machines to replicate tasks typically requiring human intelligence. Early AI approaches often focused on rule-based systems to encode knowledge and reasoning capabilities. These systems excelled in structured environments however, they faced challenges in adapting to the complexity of real-world data, particularly in areas of perception. To address these limitations, the late 20th century saw a shift toward connectionist approaches, particularly artificial neural networks, which emphasized data-driven learning. Advances like backpropagation, increasing computational power, and the proliferation of large datasets catalyzed the adoption of these methods, laying the groundwork for modern DL.

While DL demands substantial data and computational resources, it has revolutionized various fields, particularly in image analysis. Unlike traditional symbolic AI, DL models automatically extract features from data, making them particularly effective for handling data variability. CNNs are particularly pivotal in computer vision, excelling in tasks like image classification, object detection, and segmentation. Architectural milestones, including AlexNet, VGG, and ResNet, have significantly enhanced the capability of CNNs to process and interpret visual data. More recently, ViTs and ConvNext have introduced advancements, highlighting the ongoing evolution of neural network architectures.

The success of DL relies heavily on the availability of high-quality data for training. Large datasets like ImageNet have been instrumental in advancing DL models, although data augmentation techniques remain essential for enhancing model performance when datasets are limited. In the context of NDE, careful consideration is required to ensure augmentation methods preserve the integrity of defect-related data. As DL continues to evolve, its adaptability and effectiveness position it as a cornerstone technology for solving complex challenges in fields like NDE, particularly in automated analysis of ultrasonic data.

2.1.5 Artificial Intelligence in Non-Destructive Evaluation

The use of AI systems for NDE have been explored as an alternative solution to manual data processing. Historically, simple rule-based algorithms, such as thresholding or comparison to reference images were used [89], [90]. However, they were often unable to replicate human level performance due to a lack or reliable results as a result of noise and variations from operational environments [91]. As a result, human operators remain more trusted as they can adapt to changes in the data.

Despite these limitations, rule-based systems have demonstrated notable successes. For instance, commercial software tools such as “NDT kit” offers some automated detection capabilities based on pre-defined data processing steps such as thresholding or subtraction from a reference scan. This acts as an operator assistance tool, however it still requires significant operator interaction, for example to narrow down the area

of interest [92]. Despite not being a fully automated solution, analysis times have been greatly improved with such tools. Similarly, “ADA” represents an effort to encode inspection procedures into an automated data analysis pipeline [93]. While promising, ADA faces challenges in rejecting noise artifacts caused by scanning processes—artifacts that an experienced human operator would intuitively dismiss as non-defects. This approach also depends on precise parameterization of acceptance criteria, typically calibrated to reference samples. Such characteristics make it well-suited to aligning with regulatory requirements. However, these systems have demonstrated efficacy only in scenarios involving simplified geometries and have yet to prove their robustness in inspecting complex industrial components or challenging test conditions. For dealing with complexities in geometries Guo et al. suggested the use of a model-based approach, which relied on adjusting gating parameters to align with the digital representation of the component [94]. Whilst this demonstrated promise, it requires an accurate model of the component and correctly aligning or registering this model to the real-world component.

Rule-based approaches have shown utility for processing signals with well-defined characteristics. They benefit from explainability but are limited in their ability to handle variability and complexity caused by environmental changes, manufacturing inconsistencies, and intricate geometries or materials. These limitations have prevented their widespread adoption in complex industrial applications. Data-driven ML methods, by contrast, offer greater adaptability, as they are not bound by rigid predefined parameters. A meta-analysis by Sergio Cantero-Chinchilla et al. [14] underscored this gap and highlighted DL as a promising solution, owing to its capability to extract complex features from data.

2.1.5.1 Machine Learning in Non-Destructive Evaluation

The application of ML in NDE is still a relatively young field, with the earliest found example published in 1992 [95], however there are plenty of examples of researchers demonstrating the effectiveness of different methods for specific problems.

As the problem space in NDE is highly variable depending on material, geometry, defects of interest, operating conditions, and sensing modality, comparing different ML applications in NDE can be challenging and expecting models to transfer between all NDE scenarios is at this current moment not possible. This therefore makes direct comparisons in literature challenging. One method which helps to break down and compare ML in NDE literature is by grouping based upon data analysis type and material application. Such as, ML applied to timeseries data (e.g. [96], [97], [98]) like ultrasonic waveforms or to image level analysis (e.g. [99], [100], [101]), typically constructed from multiple timeseries sources. The benefit of this is that it allows for comparison of more closely aligned ML methods, as for example image level ultrasonic analysis may have more in common with image level thermography analysis than time-series A-scan analysis.

While grouping ML approaches in NDE by data type and material application aids in facilitating comparisons between similar approaches, defining a true (SOTA) remains challenging. The NDE field encompasses a wide range of inspection tasks, from defect detection to material classification, each with unique requirements influenced by component geometry, material properties, data availability, defect characteristics, and sensing modalities. This diversity complicates the development of universally applicable models and creates difficulty in establishing reliable benchmarks. Unlike fields like computer vision, where standardized datasets such as ImageNet [79] enable direct performance comparisons, NDE lacks widely accepted public datasets. This absence limits cross-study validation and makes it difficult for the community to identify and agree upon SOTA methods or approaches.

By contrast, to other fields NDE data is often restricted by industrial data-sharing and is highly specialized with formats that vary widely, from time-series data to volumetric imaging. Without accepted datasets, researchers rely primarily on proprietary data, limiting the ability to replicate findings and preventing a cohesive comparison across different studies. This reliance on specific, often inaccessible datasets restricts the establishment of a widely recognized SOTA model within the NDE community.

Despite these challenges, certain ML architectures or approaches that have achieved success elsewhere can be applied in NDE, though they typically require adaptation to meet the distinct requirements of each application. For instance, U-Net based architectures have been widely accepted as the SOTA for image segmentation [102]. The modified approaches can then be compared to the widely established SOTA for similar tasks, as was done with DefectDet [103], which was introduced as a SOTA approach for defect detection in B-scans of metal samples and compared to other commonly used object detection architectures, but its results remain unvalidated beyond the original study and application due to the inaccessibility of the dataset. The need for customization of techniques to NDE highlights the difficulty of establishing a one-size-fits-all SOTA model in NDE, as models must frequently be tailored to suit highly specific conditions. These task-specific modifications, while effective for individual cases, further complicate the cross-comparison of ML models across studies.

Some researchers have taken steps to address data availability issues by making their datasets and code publicly accessible. For example, a dataset generated from fabricated steel plates with manufactured flaws was released as an open-source ultrasonic imaging dataset USimgAIST [104], containing over 7,000 images of steel plates with

and without flaws. While this initiative represents an important step toward more open NDE ML research, such contributions remain rare and highly specific, underscoring the need for broader efforts to build shared benchmarks and encourage collaboration across the field.

Another trend in ML in NDE literature, is that older works focused on analysing time series data. These often used a large amount of hand crafted or extracted features to reduce the dimensionality of the data [96]. Whilst time-series analysis is still popular, thanks to improvements in the CV field and computing performance, a greater percentage of recent works have focused on image-based methods [99]. While ML has been applied across various NDE sensing methodologies, the literature often emphasises metal welds as a primary focus, with limited exploration of composites. Composites, which are frequently imaged differently and possess distinct properties from metals, remain relatively underexplored in this context.

For each task, regardless of data type, most applications of ML in NDE typically fall into one of three categories. Gardner et al. [7] suggested a hierarchy for automated NDE problems, summarising and explaining these problems. This was based on the Rytter's Hierarchy previously used in Structural Health Monitoring (SHM) [105]:

1. DETECTION: the method gives a qualitative indication that damage might be present in the structure.
2. CLASSIFICATION: the method gives information about the type of damage.
3. ASSESSMENT: the method gives an estimate of the extent of the damage.

The hierarchy provides a clear picture of both the importance of each task and increasing complexity. It is also important to recognise that different industrial

scenarios will have different requirements and whilst full defect assessment such as type, size and position may be required for some settings, others may just require limited detection capability. The hierarchy is in strong agreement with the automated data analysis pipeline proposed in Figure 2, which groups Classification and Assessment under “Characterisation”, whilst demonstrating their integration into a broader automation framework.

As previously discussed, ML methods rely on training data, with increasing quantities of data required often for more complex methods. The availability of training data is therefore a primary concern when applying ML to NDE. NDE encounters challenges in generating large datasets of real samples and defects, primarily due to the limited availability of such samples, particularly of defective examples, and the significant time investment required to collect the data. To address this challenge, researchers often resort to employing transformations and other techniques to augment the datasets. Alternatively, other papers have proposed using simulated data with encouraging results [100], [106]. However, for this to be effective it is crucial that the simulated data accurately reflects the problem. There is often a challenge when using synthetic data that when testing the model trained on synthetic data one must be careful not to commit an ‘inverse crime’ [107], by simply validating the results from features extracted from a synthetic model with synthetic data generated from the same synthetic model. This would produce a result that does not generalise well to the real problem. To negate this, it is therefore preferable to test models using real experimentally acquired datasets.

As highlighted by Cantero-Chinchilla et al. [14], in order for deep learning to see adoption in NDE by both industry and regulators the explainability of models is a key requirement which will help to build trust in the results. Explainable AI (XAI) is a significant, ongoing area of research which involves many different approaches to understand the reasoning behind AI systems [108], [109]. These approaches can range from model-specific to model-agnostic explainability approaches, but there is generally a trade-off between model accuracy and interpretability. Models which give a level of uncertainty in their own results is one method of helping to build trust in their outputs, as it will help to de-mystify the "black-box" of deep learning. Abdar et al. [110], have performed a comprehensive review of uncertainty quantification methods in both traditional ML and DL applications. They use the medical field as a case study for the need for uncertainty quantification. There is significant overlap between the medical and NDE industries when applying DL, for example the challenges in acquiring enough data and the conservative nature of the industries. An alternative method to explainability is using symbolic representations of expert knowledge and formalising these representations as a rule-based system. Young et al. [111] explored how these representations could be both manually and automatically captured and formalised for industrial fault detection systems.

For a detailed review of DL data analysis in NDE please refer to Cantero-Chinchilla et al. work [14]. This review paper provides a comprehensive summary of DL, automated ultrasonic methods, and how these relate to different levels of industrial automation. Whilst this work focuses on ultrasonic data, many of the conclusions and challenges identified are transferable to different sensing modalities. For example, the authors suggested different classifications for levels of automation; based upon

automation levels published by the European Union Aviation Safety Agency [112], [113]. This can be broadly applied to different NDE methodologies:

Level 0 – Classical NDE

Level 1 – Operator assistance

Level 2 – Partial automation

Level 3 – Operational automation

Level 4 – Full automation

2.1.6 Summary of Key Challenges

The automation of NDE data analysis remains a significant challenge. While efforts have been made to develop automated systems using rule-based logic, their adoption in industry has been limited. A key reason for this is their lack of flexibility and limited applicability to real-world inspection scenarios, where human interpretation often plays a critical role in interpreting data from complex geometries or handling variations from environmental changes or anomalies. ML methods offer a promising alternative by enabling systems to "learn" the flexibility required to handle these complexities—something that is difficult to encode explicitly. DL, as the SOTA in many perception tasks, including CV, presents a compelling avenue for exploration in this context, despite inherent challenges such as the need for greater explainability.

DL research in NDE is at a relatively early stage, but it has still seen varied application to different inspection scenarios. For UT this is mainly applied to metallic welds, with a distinct lack of research in the application of DL to composites. From the current literature, it is evident that there are key shared challenges when applying DL to NDE. These are outlined in the remainder of section 2.1.6.

2.1.6.1 Lack of Training Data

Lack of training data may be the biggest challenge when applying DL to NDE. This holds especially true for naturally occurring, real defects, which have not been artificially manufactured, as they provide responses that more accurately represent real-world responses. However, gathering sufficient examples of naturally occurring defects for training is challenging, and accurately establishing and labeling the ground truth for these defects presents an additional difficulty. Meanwhile, training on manufactured defects can lead to models which do not perform well on real defective data. There are two main approaches to overcome the problem of overfitting due to small datasets:

1. **Experimentally increase the dataset**: This is challenging as NDE real-flaw data is not readily available and using manufactured defects can lead to a lack of generalisation (as previously discussed).
2. **Augment the available data**: This method has been widely and successfully adopted in other ML applications, such as computer vision [88]. It has exhibited use in NDE as well in different forms. However, successfully augmenting a small dataset to cover the wide application of real-world flaws is a challenging task, which requires a well thought out methodology based on good understanding of the NDE modality and the targeted defects. This is currently the most widely adopted approach in NDE literature. Where researchers will generally create datasets from a handful of manufactured signals, then augment these to produce datasets ranging in sizes anywhere from 100's to over 1 million (typical papers use an augmented datasets on the scale of four order of magnitude) [114], [115],

[116].

In NDE scenarios, there are various methods to augment training data:

- Simple linear data augmentation or sub-sampling techniques, akin to those commonly employed in other CV applications, are often utilised for images [115].
- Virtual flaws can be introduced, where flawed signals are implanted into noisy background data to augment the number of flawed signals [116], [117].
- Different types of transfer learning can be applied to mitigate overfitting of networks when dealing with small datasets [118].
- Simulated data, such as finite element analysis (FEA) simulations, or a combination of simulated and real data, can be employed to expand the size of datasets [99].
- Generative Adversarial Networks (GANs) can be leveraged to generate entirely new synthetic datasets [119], [120].

Regardless of the technique employed, the crucial takeaway is the importance of enhancing the networks' ability to generalise by exposing it to a greater distribution of samples during training. There are a wide range of other approaches for reducing model overfitting include adjusting hyperparameters or implementing dropout etc. [121], [122]. However, it is widely acknowledged that if the initial data is of poor quality or not representative of the target domain or distribution, the model performance will invariably be subpar.

2.1.6.2 *Defining the Problem Scope*

When implementing a DL methodology, it is crucial to have a well-defined problem. This ensures that effective training data, representative of the problem, can be established. When applying DL to NDE, the problem can often be broken down into; what is the task (detection, classification, assessment etc), and what is the target domain (signals, scans, materials, defects etc.). For example, characterising a defect in steel welds is different to detecting a defect in composites, which is a different task to segmenting cracks in concrete, and the training data will need to reflect this. When trying to detect defects it is important to not only understand what types of defects are likely to occur, but also to understand what types of signals and scans are most appropriate. This dictates what will serve as the model input. All these factors influence the applicability of different types of models. Depending on the specific scenario, it may be appropriate to construct a custom network, while in other cases, utilising an off-the-shelf model and fine-tuning it to meet the task requirements may be more suitable.

2.1.6.3 *Model Evaluation and Explainability*

NDE is frequently employed in highly safety-critical industries. However, the adoption of new technologies in these sectors tends to be conservative, given the paramount importance placed on safety and reliability. For adoptions of DL/ML methods into these industries, it is important to effectively communicate and evaluate the performance of these models. A great way of doing this is by comparing them to human operators and POD scores, as well as traditional ML statistical evaluation methods. The testing of models should also be done very carefully with a large enough,

unseen, test set that is representative of the target domain. This should also be pushed past the target domain to determine the models' limitations. Furthermore, it is essential to demystify the 'black box' nature of deep learning DL, particularly in highly regulated industries. To achieve this, XAI methods should be employed to make it easier to understand how models arrive at their predictions and what influences specific decisions. XAI can enhance transparency and is likely to be crucial for meeting regulatory standards and ethical requirements, helping to build trust and allowing DL models to be safely integrated into critical applications such as aerospace. Researchers should also be clear about the limitations of models even if these are a result of the NDE method itself.

2.2 Experimental Data Collection

Much of the work in this thesis was conducted using UT data collected experimentally with a robotic system from composite samples. The methodology for data acquisition is largely consistent across the chapters of this thesis. Hence, this section provides background information about the data collection process and the samples used.

The ultrasonic data was acquired at room temperature using a robotically deployed unfocused linear phased array. The array used was an Olympus Inspection Solutions RollerFORM-5L64 [123] (Figure 15), which had a central frequency of 5 MHz and was made up of 64-elements with a pitch of 0.8 mm and elevation of 6.4 mm. The roller probe was initially filled with deionised water as per the manufacturer's guidelines; however, this was later switched to non-corrosive glycol. Glycol has similar acoustic properties to water and is an appropriate alternative as per the manufacturer's guidance. The benefits of using glycol were two-fold. Firstly, when using water, the array was required to be removed and cleaned regularly, as constant submersion led to an increased risk of delamination of the array's protective layers. Secondly, the increased viscosity of glycol meant that the formation of microbubbles on the surface of the array was less common which minimised any acoustic barriers. The elements were driven at 100 V with a receiver gain of 22.5 dB to maximise the front wall amplitude without signal saturation. The sample rate was 100 MHz. A Peak NDT Ltd. MicroPulse 6 or LPTA [124] (both supporting at least 64 transmit/receive channels) was used for ultrasonic control with a digital band pass filter applied on reception to filter out frequencies between 2 and 6 MHz. A summary of roller probe parameters is given in Table 2.

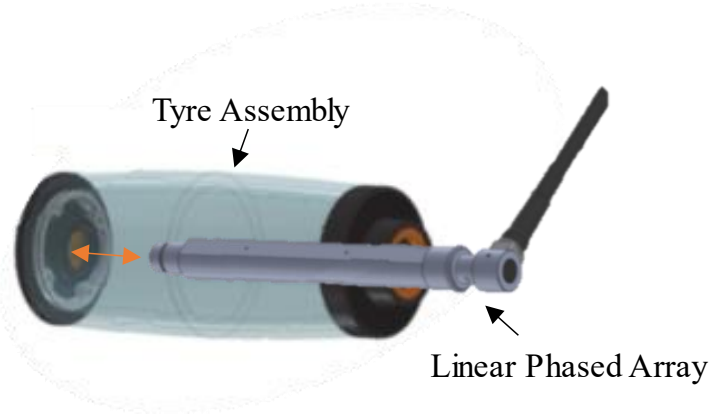


Figure 15: Diagram of roller probe assembly. Adapted from [124].

Table 2: Summary of roller probe parameters

Manufacturer	Olympus Inspection Solutions
Model	RollerFORM-5L-64
Central operating frequency	5 MHz
Number of elements	64
Pitch	0.8 mm
Elevation	6.4 mm
Driving Voltage	100 V
Receiver Gain	22.5 dB

The pulse repetition frequency was set to collect a 4-element aperture, unfocused B-scan every 0.8 mm with A-scan speed of 10 mm/s which was controlled using a fully automated robotic system built around a KUKA KR 90 R3100 extra HA industrial robot (Figure *17*) [125]. Robotic scanning enabled the concatenation of encoded B-scans to form volumetric datasets. Robotic scanning was essential in accurate concatenation of scans. This process was highly repeatable thanks to a ± 0.04 mm pose repeatability. To ensure a steady coupling of the roller-probe to the surface of the component and consistent transfer of acoustic wave energy into the sample at different scanning positions, Force-Torque compensation was used to control the contact force on the samples surface with feedback from the force axis perpendicular to the sample. This was accomplished with integration of a Schunk GmbH & Co. FTN-GAMMA-

IP65 SI-130-10 Force-Torque sensor, mounted between the robot's flange and the roller-probe. Initial experiments used a contact force of 35 N (Chapter 3), however latter experiments increased the force to 70N which increased the consistency of acoustic transmission by ensuring consistent tyre compression throughout the scan. The Force-Torque sensor had a programmed limit of 130N in the primary force axis. This was set to avoid any potential damage caused by errors during the inspection. Water was used as an acoustic couplant in the scanning process. Similar data acquisition setups are used in industry and has been employed for data collection on large composite aerospace components previously [26].

Communication between the robotic interface and data acquisition was handled through a custom LabVIEW interface. A custom binary file format (.rput) was developed for this use case to efficiently store ultrasonic, robotic, and Force-Torque data for each B-scan. The development of the binary file format resulted in an approximate 2 times reduction in memory compared to previously employed CSV file formats. The robotic information included the position at each frame which allowed for accurate rasterisation of multiple scan passes. An accompanying Python script decoded the custom binary file into usable NumPy arrays which allowed for efficient post processing and ML development in Python. A block diagram of the setup is illustrated in Figure 16.

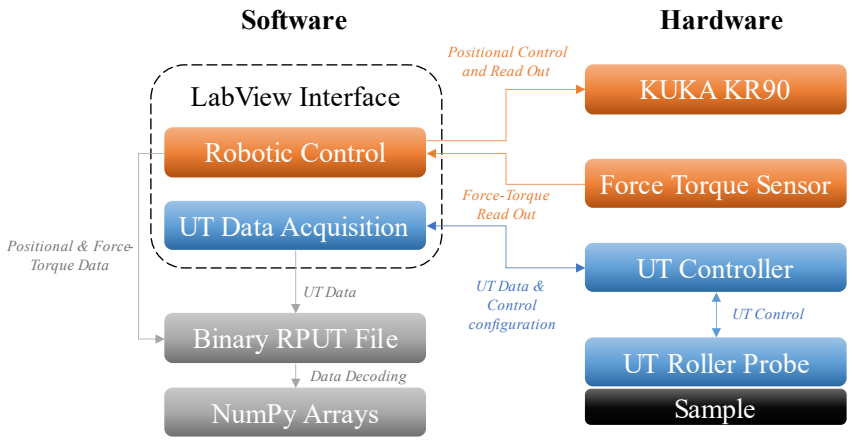


Figure 16: Block Diagram illustration of the experimental data acquisition setup.

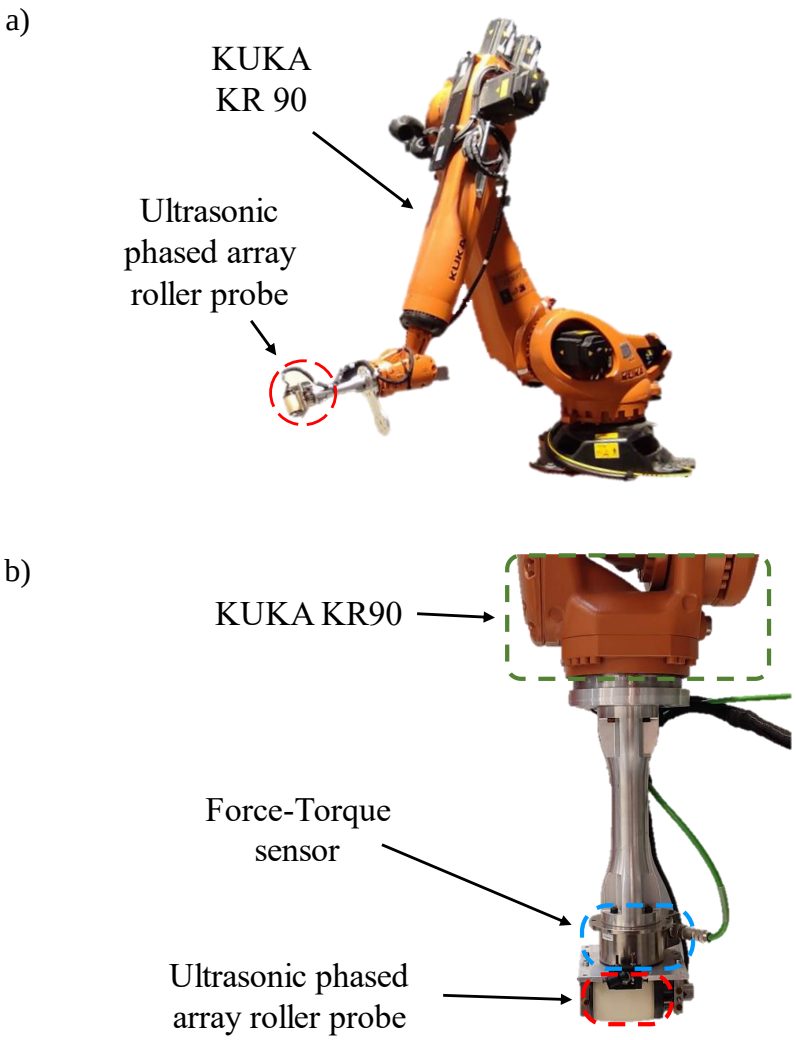


Figure 17: a) Overview of the experimental setup of KUKA KR90 and ultrasonic roller probe used for data acquisition. b) Close-up image of the experimental setup showing the assembly of the roller-probe and Force-Torque sensor as the robot end effector.

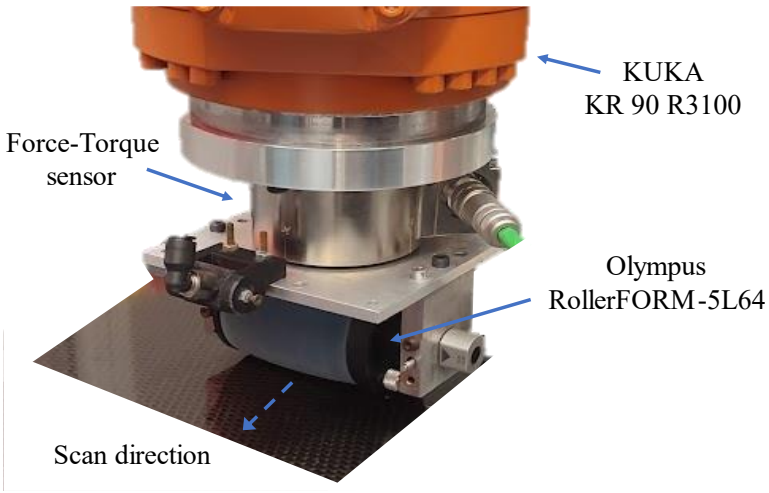


Figure 18: Overview of the experimental scan setup of KUKA KR90, Force-Torque sensor, and ultrasonic roller probe used for data acquisition.

Experimental ultrasonic data was acquired from CFRP samples, both with and without artificially introduced defects, to serve as test data. To imitate delamination defects, which are one of the most common defects in composites [101] and a significant life-limiting failure mode [126], Flat-Bottom Holes (FBHs) were drilled from the backside of the samples. Such defects are simple to produce post-cure and give similar responses to delamination's. In addition to this, their consistent known geometry allows for them to be used as references for defect sizing.

Prior to introducing defects, clean scans of each sample were taken to form defect-free datasets. The use of the same CFRP base sample ensured that the trained models learned defect-specific responses rather than the underlying properties of different composite samples. Two 254 x 254 x 8.6 mm (WxDxH) composite samples were provided by Spirit AeroSystems. The samples were all manufactured to the BAPS 260 specification using a Resin Transfer Infusion Process, made using non-crimp fabric and Cycom 890 resin. The ply layers had a repeating lay-up pattern of 0, 45, -45, and 90 degrees and a density of 1440 kg/m^3. In the first sample 15 FBHs were drilled from

the backside to simulate defects. The defects were 3.0, 6.0 and 9.0 mm in diameter, with each individual defect diameter drilled to depths of 1.5, 3.0, 4.5, 6.0, 7.5 mm from the front surface. The different defect diameters were spaced 30 mm apart with different depth defects separated by 35 mm. In the second sample, 25 FBH were drilled to the same depths as the first sample but with additional defect sizes of 4.0 and 7.0 mm as shown in Figure *19*. All defects were manufactured to tolerances in depth of +/- 0.3 mm, and diameter of +/- 0.2 mm. A summary of the samples is provided in Table 3.

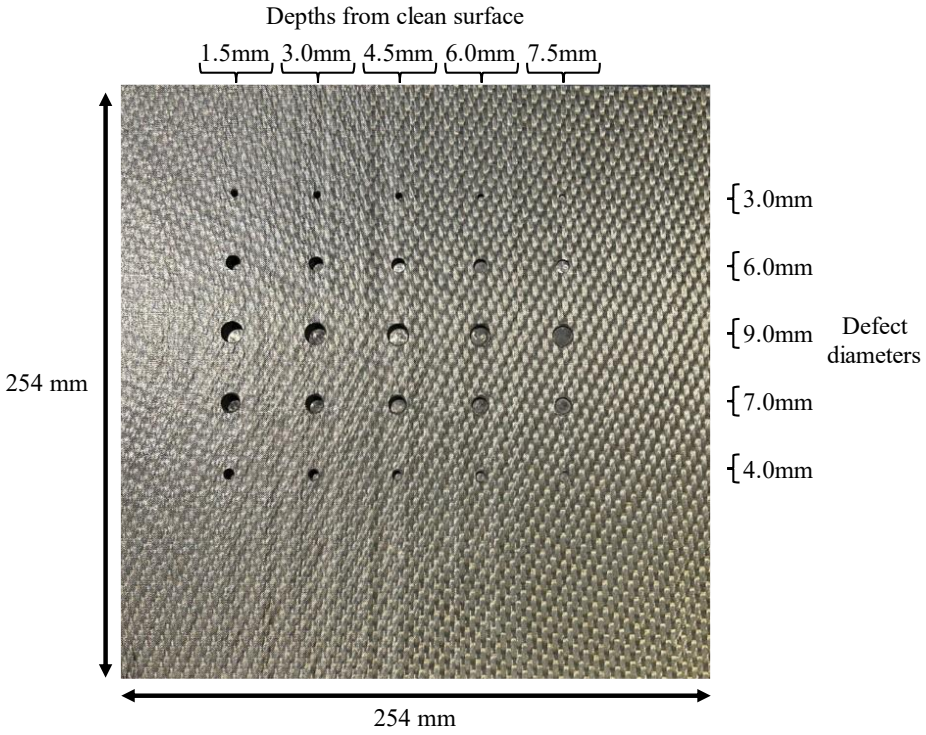


Figure 19: The composite test sample showing 25 FBHs.

Table 3: Summary of samples and their defects.

Sample	Number of Defects	Diameters (mm +/- 0.2)
1	15 Flat-Bottom Holes	3.0, 6.0, 9.0
2	25 Flat-Bottom Holes	3.0, 4.0, 6.0, 7.0, 9.0

2.3 Signal Processing

There are many different signal processing techniques that can be applied used to increase the interpretability of the raw radio-frequency time series data collected from each transducer (or group of transducers). Commonly when imaging, the envelope of the signal is taken, this is often done using the Hilbert transform. The Hilbert transform is used to obtain the analytical signal, the absolute of which gives an enveloped signal which is useful for calculating the instantaneous response of a time series. This is beneficial when imaging as it not only gives a positive response, but it also smooths the signal, unlike simply rectifying the signal, where the absolute value of response is taken. This approach is a standard signal processing technique used when generating C-scan images from time series ultrasonic data [127]. The theory and mathematics behind the Hilbert Transform are well documented, for further details please refer to the relevant literature [128].

In many UT scenarios, the signals obtained yield strong responses from geometric features, such as the front or back wall response. Including these features in the imaging process can often mask out other useful information due to their high amplitude response. To address this issue, gating is frequently employed in the time domain to selectively remove these geometric features and establish a window that focuses the imaging specifically on the area of interest.

Many different signal processing methods are used in UT. Examples of these include the Wavelet transform; where a signal is decomposed into shifted and scaled versions of a base function (or wavelet), the Fourier transform; where the signal is decomposed into sine waves of specific frequencies, signal filtering; where a portion of the signal

is removed, this is often done in the frequency domain but can also be done in the time domain to remove features not of interest, or enveloping; where the boundary of the oscillating signal is extracted. The use of different pre-processing techniques is often application dependent, for specific details on these please refer to the relevant literature [127], [129], [130]. The work conducted in this thesis primarily involves the use of the methods detailed in the following section.

For the work conducted in this thesis, data collected (either experimentally or through simulations) was in the form of radio frequency A-scans, also known as amplitude scans. To make the raw radio frequency data more useful, different pre-processing steps were used. Initially, the A-scans were centred at zero mean amplitude and enveloped using the Hilbert transform previously discussed, as shown in Figure 20 (a). Subsequently, each dataset was normalised between 0 and 1 by dividing by its maximum absolute peak amplitude. Normalisation is not only a beneficial step in data processing for ML training stability, but it also allowed for direct comparison of the different datasets as amplitudes from the simulations are relative to each other and are not reflective of experimental voltage values. In addition to facilitating dataset comparability, the data preprocessing steps enabled the incorporation of domain-knowledge into the analysis pipeline. For instance, prior understanding dictates that the front wall response corresponds to the surface of the component and, under standard inspection conditions, should consistently appear at the same point in time. While machine learning models could theoretically learn this information given a sufficiently large and representative dataset, embedding this knowledge during preprocessing helps to eliminate noise in the data and assists the models in learning effectively.

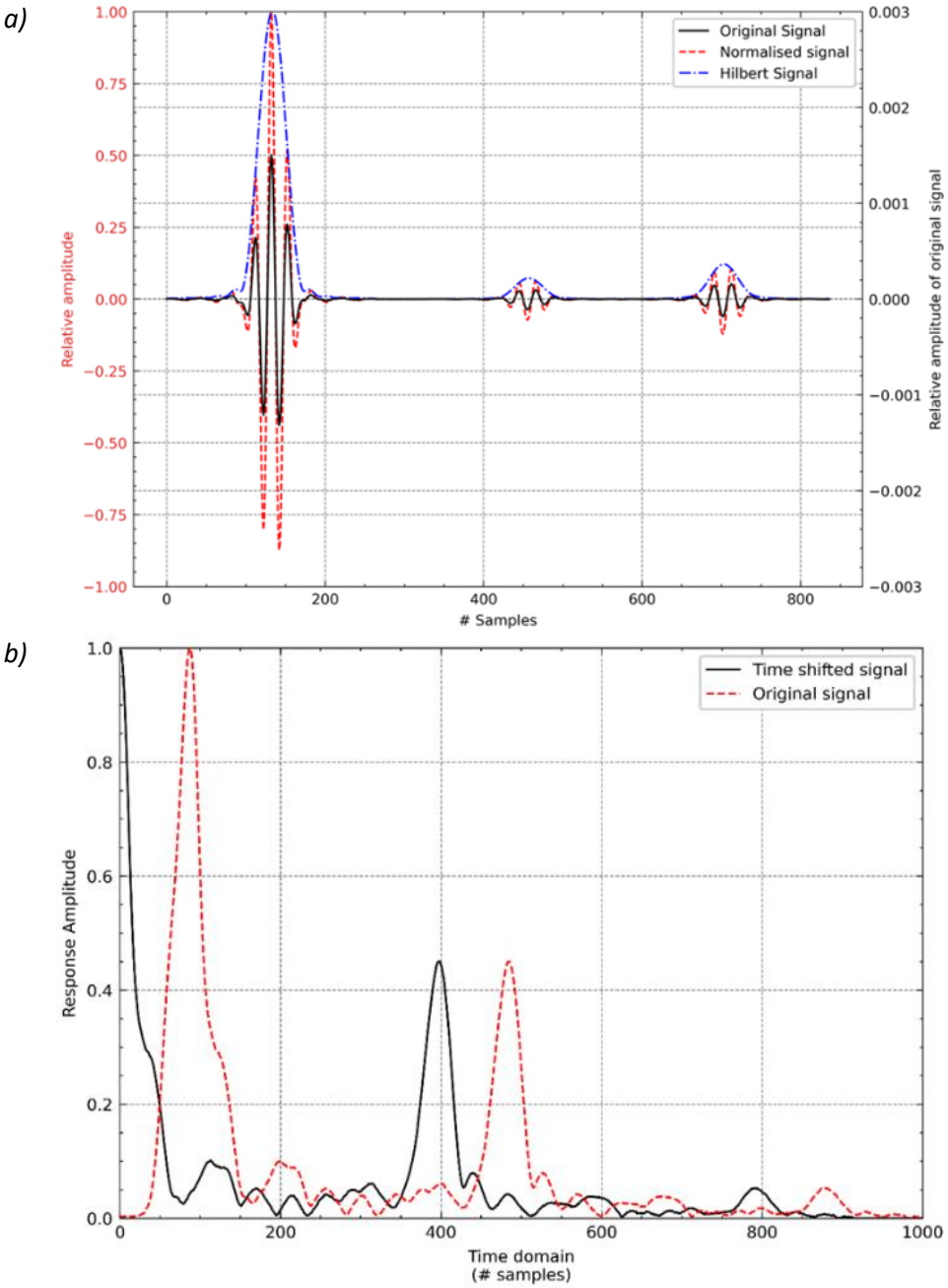


Figure 20: a) Example of relative amplitude response from simulations, normalised signal, and Hilbert transform, applied to the original signal. b) Demonstration of how individual A-scans are time shifted to the front wall response.

For chapter 3, which is focused on the interpretation of C-scan images, the front and back wall were gated out. For chapters 4 and 5, which focus on volumetric analysis, once the data was normalised, the offset in the time domain was compensated for by

aligning the peak front wall response to the origin. This made sure that features were correctly aligned in the time domain and helped to account for any variability in the acoustic path length between individual transducers and the surface of the sample. Figure 20 (b) Shows how the time shifting was done for an individual A-scan with the Hilbert transform applied. Figure 21 shows the effect of this on a complete ultrasonic experimental volume.

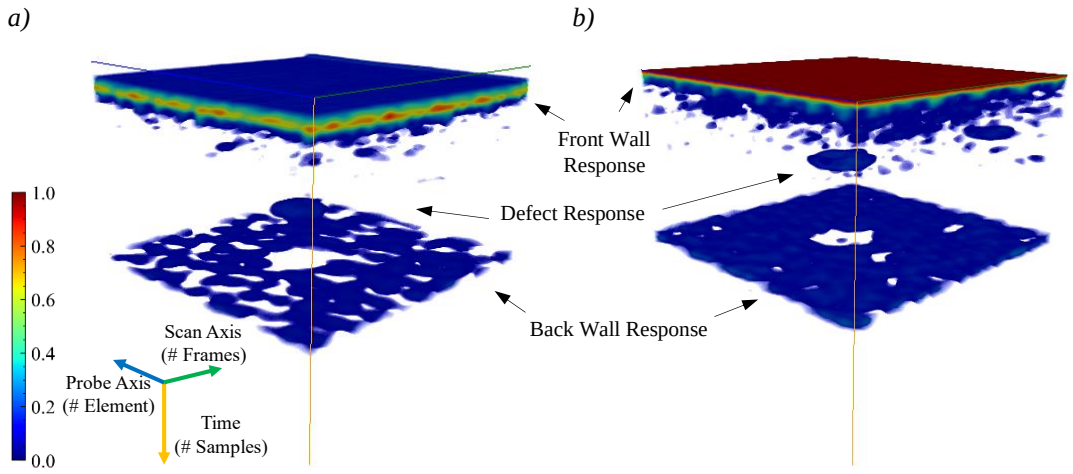


Figure 21: a) Volumetric data with Hilbert transform applied only. b) Volumetric data with time shifting to the central response of the front wall peak. Both figures have been thresholded to remove the lowest 10% of amplitudes to aid in visual clarity.

2.4 Dataset Introduction

The resolution of the UT data in the array dimension was constrained by the element pitch, and the scan width was restricted by the number of elements in the array. This limited the inspection data to 64 voxels in the array dimension. To match this, 64 B-scans were concatenated in the scan dimension to create cuboidal datasets, as seen in Figure 21. As the element pitch was 0.8 mm, and the robotic-scanning speed was regulated with the pulse repetition frequency to ensure a B-scan offset of 0.8 mm. This enabled the generation of volumes with square voxels in the spatial domains, along both the probe and scan directions. By utilising this approach, the work was able to achieve a standardised volumetric resolution that was consistent throughout the datasets. Since ultrasound values are just echo amplitude responses received from within the inspected component by the array and presented in levels of voltage response, the images/volumes were kept in single channel grayscale as any colours did not have any physical significance.

2.5 Conclusion

This chapter provided an overview of ultrasonic technology, covering topics from basic acoustic propagation to phased arrays and mechanized scanning. It also underscored the growing use of composite materials in aerospace applications and safety-critical components, which heightens the need for effective NDE to identify a variety of potential manufacturing defects. UT is widely adopted in the industry for inspecting aerospace composites due to its ease of use and reliable detection capabilities across various defect types.

The chapter also introduced foundational concepts in AI, with an emphasis on data-driven approaches like DL. A summary of the applications of these methods to NDE was presented, along with a discussion of the specific challenges encountered in integrating these techniques within the field.

Finally, the chapter outlined the experimental data collection methodology, and the samples used throughout this thesis, detailing aspects such as signal processing and dataset construction.

3 Synthetic Data: A Solution to Training Data Scarcity

3.1 Introduction

Despite the clear opportunity, ML has seen limited uptake in UT signal analysis, particularly for composite components, which present a more challenging case with additional structural noise compared to isotropic and homogeneous materials. A clear barrier to research developments is the lack of training data [14]. This combined with industrial questions over interpretability and compliance with standards has presented challenges for the use of ML, and DL particularly. Modern manufacturing processes aim to reduce the production of defects, meaning large volumes of real defect responses are simply not available; especially ones that represent the full distribution of defect classes and wide variability within these classes that are present from inconsistencies in manufacturing. Furthermore, the manufacturing volumes of aerospace components can be small, and stringent protocols for data protection of civil and military components compounds the issue of data scarcity. Most commonly, previous works have aimed to experimentally increase their datasets using manufactured defects [32], [131], [132]. However, whilst these approaches can demonstrate research concepts, they are unlikely to give UT responses that accurately represent real-world responses especially not at the same variability seen within real defects. Other authors have demonstrated success using simulated data developed using FEA software to model defects and ray-based models to create Plane Wave Capture, which uses a physics-based understanding of the wave propagation to produce accurate responses based on bulk material properties [106]. However, this is typically done for isotropic and homogenous steel samples which have very low attenuation and noise, and have less modelling complexity compared to composites,

which are acoustically anisotropic and produce large amounts of UT wave attenuation and scattering noise. Furthermore, this noise is often produced structurally from the internal ply/fibre bundle interfaces of the composite material and is not random. Therefore, neglecting this structural noise component and merely addressing the random noise through the addition of randomly distributed noise to the datasets may give unrealistic images or obscure defect responses. It is therefore important to understand what gives rise to the complete noise, how this can be modelled, and how this impacts our DL models. Most modern FEA software can account for ply interactions, but it needs intensive material acoustic property characterisations, modelling effort, and very long time-transient simulations. Therefore, composites are often modelled using average bulk properties and not done at the individual ply level. As an alternative to full FEA software, semi-analytical physics-based software has been shown to produce experimentally accurate defect responses [133], [134]. This software is much less computationally expensive than full FEA and can be used for simulating composite responses based on bulk material properties [135]. This provides a great opportunity to simulate vast amounts of defect responses with low computational cost however, it does lack the complexities of structural noise response.

Synthetic datasets are widely used in ML to augment small training datasets [136] and they offer a potential solution to the lack of defect data in UT. This chapter looks at different novel methods of generating synthetic datasets from simulated data for composite UT. These novel synthetic data generation methods are comparatively evaluated on their experimental classification performance when used for training a CNN. Hyperparameter Optimisation (HPO) is used to select an appropriate CNN architecture that can represent the solution space for our task. GAN [137] are one of

the approaches investigated and have seen success in generating and augmenting training data.

GANs are a class of generative ML models which rely on the relationship between two networks: a generator and a discriminator. During training the two networks learn together; the generator tries to produce data that the discriminator is unable to distinguish from the ground truth, whilst the discriminator learns to better identify real and generated data. GANs are often used to create or fill in images and as such have been explored as a means for augmenting limited data sets for CNNs. To augment data, GANs are often used to generate additional samples to further populate a distribution of a particular target case, relying on the variability within the GAN to provide a greater variability in training examples [119], [138], [139]. An additional benefit of using synthetic data generated from GANs could be in the anonymisation of the original dataset to address security or privacy concerns [140]. This could help to alleviate concerns over data sharing of real NDE defects from industry.

The specific GAN used in this work to tackle a data shortage challenge for the first time in the NDT domain is CycleGAN, which is a conditional GAN that has demonstrated good results in unpaired image-to-image translation tasks [141]. This GAN approach aims to combine NDT data generated from physics-based simulations with GAN augmentation to create a dataset based upon physically accurate defect responses that better resemble experimental data. The approach uses a modified CycleGAN architecture to learn the mapping from simulated UT data to experimental UT data. Specific, novel modifications, integrally an additional loss function, help to encourage accurate defect signal reproduction whilst allowing for the addition of

experimental noise. With this approach, large quantities of highly varied simulated defects can be produced in a relatively short time as compared to experiments or FEA, and using the GANs mapping, produce large quantities of experimentally representative synthetic data. The overall goal of the work in this chapter is to identify the best methods for generating synthetic datasets in UT of composites to help unlock the potential of DL in NDT applications.

This chapter provides details on how experimental and simulated UT testing is gathered and processed into defective and non-defective image datasets (section 3.3). In section 3.4, information is provided on the use of a CNN architecture for evaluation of classification performance and details on the HPO method used for architecture selection. Comparison is made with the experimental classification performance between simulated and experimental data in section 3.5. The different methods of synthetic data generation are then explored in section 3.6 with the effects on classification performance evaluated. Section 3.7.1 provides a summary of the classification results. Finally, section 3.7.2 introduces Grad-CAM as a method to help with model interpretability when comparing synthetic data to experimental data and discusses the full results of this work.

3.2 Simulated Data

A simulated dataset of the experimental test sample discussed in section 2.2 was constructed using a semi-analytical, physics-based, commercial NDT simulation software – CIVA [142]. Flaw interaction in CIVA is made up of three computation stages: incoming transient ultrasonic field arising on the defect, field-to-flaw interaction according to the Kirchhoff approximation, and prediction of the sensitivity

at reception using Auld's reciprocity theorem. The Kirchhoff approximation assumes the wave does not propagate into the defect which is appropriate for the FBHs modelled here [133], [143]. As the software adopts a semi-analytical approach, it allowed for simulations to be completed with significantly reduced computational cost (with an approximate 60x reduction in time) compared to FEA methods. Since the focus of this work was the opportunity to produce large datasets for UT, this was a significant benefit of the semi-analytical software approach.

CIVA simulation software is physics based, and has been widely used for commercial UT simulation work, and experimentally validated for UT [133]. Therefore, there was reasonable confidence that the modelling of wave propagation and its interaction with defects were representative, producing reliable defect responses as well as being computationally efficient. In addition, the simulated defect dimensions and positions were readily controlled, allowing duplication of the exact experimental setup. This allowed for efficient, complete annotations of the dataset to be generated at the point of simulation, which opens further opportunities beyond classification, such as segmentation where each pixel within the image is individually classified which could be beneficial for defect sizing etc. A significant downside of using a semi-analytical software as opposed to FEA is that the software was unable to model each distinct composite layer response leading to differences between the simulations and measured experimental responses, such as the lack of coherent scatter from ply interfaces. However, in creation of the model, the individual layers were still constructed but were only used to estimate equivalent homogeneous (and anisotropic) material properties. A single ply layer was constructed and alternated with 0, 45, -45, and 90 degrees to match the experimental sample as closely as possible. The resulting multilayer

structure was homogenised so that it was consistent with a homogeneous medium having mechanical properties equivalent to those of the multi-ply composite. The homogenisation gives good bulk propagation characteristics but removes structural noise due to reflections from ply boundaries. This is a limitation of this simulation method but is a necessary trade off against the computational cost of FEA when producing large datasets. The fibre density was also set to 50 % to give the density which best matched the experimental sample value of 1440 kg/m^3. A parametric study simulation was setup which used the composite bulk properties previously calculated and varied the diameter and depth of defects. The study matched the experimental setup with 3.0, 6.0 and 9.0 mm defects at depths of 1.5, 3.0, 4.5, 6.0, and 7.5 mm from the surface. Both the front and backwall surface reflections were included in the model. The full simulations for the complete dataset took less than 6 hours on a desktop computer with a 24-Core 3.79 GHz CPU and 128 Gb of memory.

3.3 Image Generation

Once the data was normalised and the signal processing steps outlined in section 2.3 were applied. The data was truncated to remove the front and back wall echoes across the full dataset. Then the maximum amplitudes were taken at varying depths of 5 samples in the time domain to produce C-scans (sampling rate of 100 MHz), refer to Figure 8 (*C-scan depth gating*) for visualisation of the image extraction process. This enabled direct comparison for multiple different response images to be generated for each defect. From these C-scans, the images which represented a defect response were collected. For the experimental samples, data was also collected from the reference sample to obtain defect free C-scan images. In total this produced 334 defective images from the experimental training sample, 150 defective images from the

experimental test sample and 640 defect free images from the reference sample. This was split into 334 clean training images and the rest were used for testing. From the simulated dataset, 154 defective images were produced. Figure 22 shows how the simulated responses were significantly different from the experimental data. The simulated responses have far greater signal to noise ratio than the experimental responses and lacked the background response that is typically seen in experimental scans from the composite ply interactions, with a mean signal to noise ratio of over 400 times the simulated defective datasets compared to the defective test dataset. A summary of the datasets generated from the experimental and simulated data is given in Table 4. For details on experimental samples and data acquisition, please refer to in section 2.2.

Table 4: Summary of the datasets produced.

Data source	Dataset	Number of images
Experimental test sample (15 Flat-Bottom Holes)	Defective test	150
Experimental train sample (25 Flat-Bottom Holes)	Defective train	334
Experimental reference sample	Clean test	148
	Clean train	334
Simulated experimental test sample (15 Flat-Bottom Holes)	Simulated defective	154

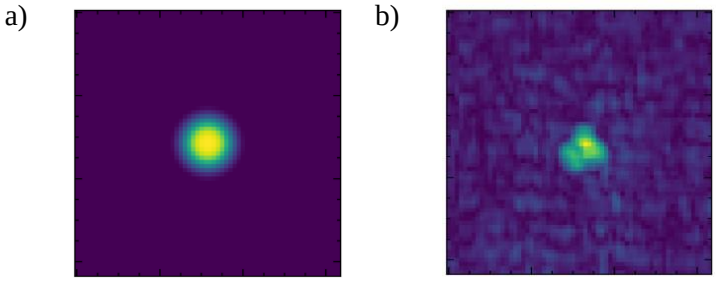


Figure 22: example of simulated (a) and experimental (b) C-scan responses of a 9 mm diameter FBHs.

3.4 CNN Classification and Evaluation

3.4.1 Classification Evaluation with CNNs

The aim of this work is to evaluate different methods of modifying simulated data to make them more effective at training Deep Learning models for experimental classification tasks. It is therefore important that we evaluate our synthetic datasets with respect to a classification metric. A CNN was used to evaluate and compare the classification performance of different synthetic and experimental datasets. CNNs have repeatedly demonstrated wide scale success in image classification and are appropriate for this task [74].

Since the focus of this work was to compare synthetic datasets and not on optimal classification accuracy, the CNN was kept constant for each dataset. Whilst the CNN should be kept lightweight to reduce the computational cost of testing each synthetic dataset, it was also important that the CNN had adequate complexity to learn the task. To make sure the CNN had enough complexity to represent the solution space, a genetic algorithm was deployed for hyperparameter optimisation (HPO) of a CNN when trained on experimental data. A genetic algorithm is a heuristic search method that mimics natural selection seen in biological evolution.

As the datasets used in the study are small, there was a degree of variability in the classification results. To negate this, when training the classifier, the CNN was re-trained for each synthetic dataset with a fresh initialisation 100 times and the average results were taken. Each CNN was evaluated on the same experimental dataset of 298 images, made up from the experimental clean and defective test dataset. Figure 23 shows the methodology used for classification evaluation.

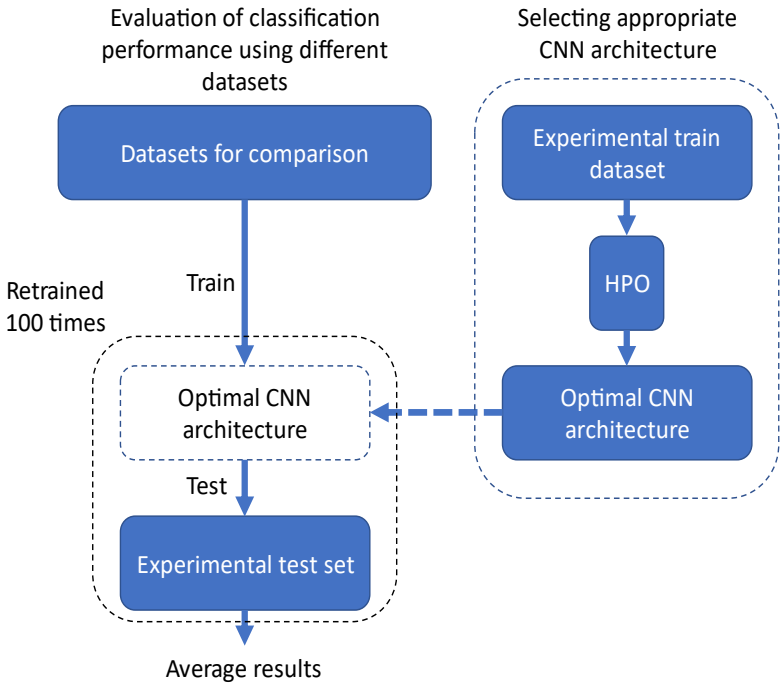


Figure 23: Flow diagram showing the process used for HPO of the CNN architecture and the use of the optimal architecture for classification evaluation.

To quantitatively assess the performance of the classifiers, confusion matrices were generated, and precision, recall and F1 scores were calculated according to (7), (8) and (9).

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

$$F1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad (8)$$

Where TP is true positive, FP is false positive, and FN is false negative, with positives being the presence of a defect. Each result was individually averaged using a simple mean across the 100 training cycles.

3.4.2 Hyperparameter Optimisation on Experimental Data

A genetic algorithm was used to perform HPO on the experimental training (defective and clean) dataset to determine the parameters for the CNN. The model had at least 1 convolutional layer. Each convolutional layer had a fixed kernel size of 3 and used a ReLU as the activation function [144] followed by max pooling with a kernel size of 2. The number of convolutional layers was parameterised with the number of filters given by a constant out-channel ratio and the number of out channels from the previous layer. The out-channel ratio was also parameterised. The network always had at least one fully connected layer, from the flattened layer to the single output node, with a sigmoid activation function for binary classification. There were a variable number of fully connected layers and each hidden fully connected layer used ReLU activation. The number of nodes on each hidden layer was equally distributed by dividing the number of nodes in the flattened layer by the total number of layers and removing this from the previous hidden layer each time. The optimised hyperparameters also included batch size, early stop, learning rate, momentum, and number of epochs. The

values for the HPO variables are given in Table 5. Figure 24 shows an example of the network with three convolutional layers, 2 hidden layers, and an out-filter ratio of 2.

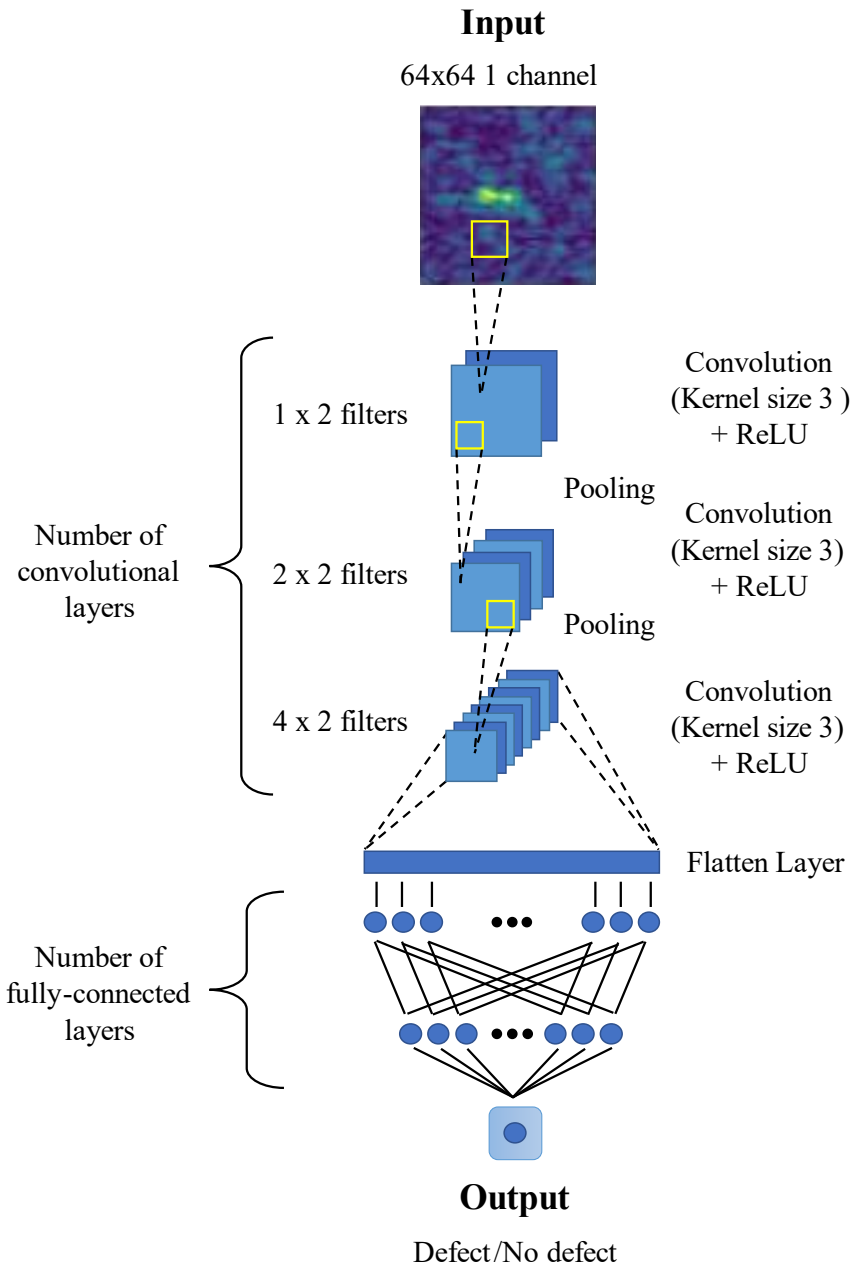


Figure 24: CNN architecture example with a convolutional channel ratio of 2.

Table 5: HPO variables and their range of values.

Variable Parameter	Range
Number of fully connected layers	1 - 6
Number of convolutional layers	1 - 6
Channel ratio for convolutional layer filters	1 - 3
Batch size	16, 32, 64, 128, 256
Early stop	0 - 5
Learning rate	0.00001 – 0.5 (log scale)
Momentum	0 - 1
Number of epochs	100 - 500

The HPO was performed using the experimental train dataset, made up of 334 defect images and the same number of defect free images from the clean train dataset. The genetic algorithm used was a variant of Regularised Evolution (RE) [145], which was adapted for continuous and integer valued hyperparameters. The algorithm was initialised with a Population (P) of 128 configurations generated via a random search. At each iteration RE sampled 5 configurations from the population, the model with the highest evaluation score within this sample was selected and a new child configuration was generated by mutating one of the parents hyperparameters. This child model is then trained and prepended to the population with the 'oldest' model discarded. This assisted in avoiding the system becoming trapped in local minima, as high performing models relative to the population will be exploited for P iterations before being discarded and allowing the process to explore new areas of the search space. This method was run for 512 iterations. During each model evaluation, the dataset was randomly subsampled without replacement with 80% of the dataset used for training and 20% used for testing. The F1 score was calculated over 10 iterations of training and testing data samples with the average F1 score used as the evaluation metric. The optimum final network had an average F1 score of 0.978. The optimum

hyperparameters are outlined in Table 6. The network was implemented using the PyTorch framework [146].

Table 6: Optimised hyperparameters used for CNN.

Variable Parameter	Optimal value
Number of fully connected layers	1
Number of convolutional layers	3
Channel ratio for convolutional layer filters	3
Batch size	16
Early stop	1
Learning rate	0.014
Momentum	0.176
Number of epochs	264

3.5 Experimental and Simulated Data Classification Performance

3.5.1 Experimental Results

For comparison to the synthetic datasets, a model was trained on the experimental test dataset and the same number of clean images sampled from the clean test dataset with a train/test split of 80% and 20% respectively. This gave a total of 60 test images. The averaged results across 100 training iterations, gave a mean model accuracy (± σ) of 89.8 ± 9.8%, with average F1, precision and recall scores of 0.887 ± 0.112, 0.974 ± 0.135 and 0.826 ± 0.119, respectively. The average confusion matrix for the experimentally trained model is given in Table 7.

Table 7: Average confusion matrix across 100 training iterations for a CNN trained on experimental data.

		Predicted	
		Defect	**No defect**
True	**Defect**	30.0	1.0
	No defect	5.1	23.9

3.5.2 Simulated Results

A model was also trained on the simulated, unmodified defect response data and the same real defect free images generated from the defective test sample which were used for the experimental results. This was made up of 154 simulated defect images and 154 real defect free images sampled from the clean train dataset. After 100 training iterations, the model gave an average accuracy (± σ) of 62.8 ± 4.9%, with average F1, precision and recall scores of 0.394 ± 0.109, 1.00 ± 0.00 and 0.252 ± 0.100, respectively. The average confusion matrix for the model trained on simulated data is given in Table *8*.

Table 8: Average confusion matrix across 100 training iterations for a CNN trained on simulated data.

		Predicted	
		Defect	**No defect**
True	**Defect**	150.0	0.0
	No defect	119.7	37.3

3.6 Methods of Synthetic Noise Generation

In this work four separate methods were explored to map simulated data to more experimentally representative synthetic datasets by adding noise. The first approach uses a modified CycleGAN to learn the mapping between simulated and experimental data. The second approach aims to utilise the fact that clean ultrasonic images are comparatively much more available than defect data, by combining both real clean images and defect simulations. The final two approaches studied the noise profiles seen in the experimental data and attempted to simulate these at both the C-scan image level and the individual A-scan level.

3.6.1 Approach 1: Simulated to Experimental Domain Mapping with CycleGAN

To learn the mapping between simulated and experimental data, an image-to-image translation GAN was used. CycleGAN was chosen as it has shown promising results in unpaired image-to-image translation, and works particularly well for style transfer tasks which this application is similar to [141]. Unlike conventional GANs that rely on adversarial loss alone, CycleGAN introduces an additional cyclic loss, which enforces reconstruction for a full cycle; from source domain to target domain and crucially back to source domain. The addition of cyclic loss removes the necessity for paired training data, allowing for unsupervised domain transfer. Not requiring paired images in training was a significant advantage as it provided greater freedom in the images used in training. Furthermore, from an NDT perspective, if this approach was extended to naturally occurring defects, it would be impossible to accurately simulate the complexity of naturally occurring experimental defect responses to produce a completely paired dataset.

Implementing the standard CycleGAN directly with the parameters given in the original paper[141], was unable to accurately reproduce ultrasonic images with the simulated defect responses present. Furthermore, the generated images suffered from significant mode collapse. Mode collapse occurs when the generator repeatedly outputs a single type of image, due to finding one image that is successful in fooling the discriminator. Figure 25 shows an example of this, where different input simulated defect responses produce the same output. The original implementation was done in Pytorch and was trained for 200 epochs, with a batch size of 4, 6 residual blocks, and an identity loss of 5 (half the cycle consistency loss).

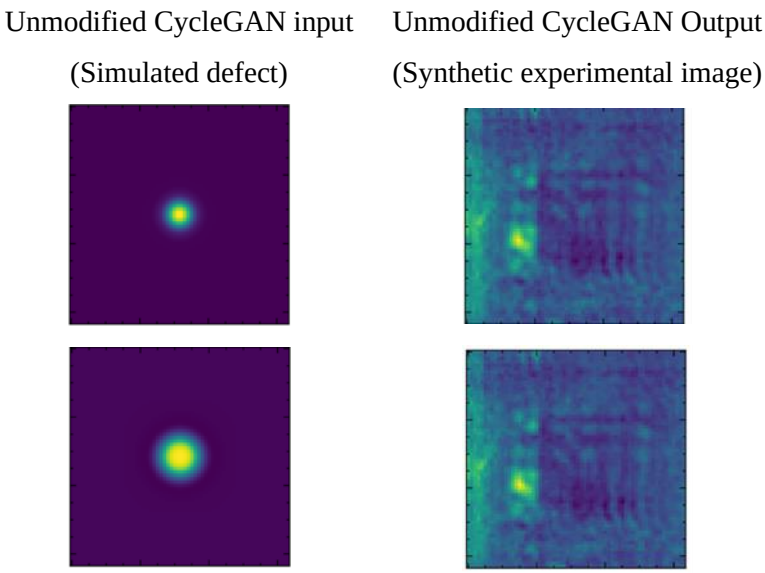


Figure 25: Example images of initial CycleGAN outputs.

3.6.1.1 CycleGAN Modifications– Mid-Cycle Activation Map

It has been demonstrated that adjusting the loss function of CycleGAN can improve performance for specific tasks [147]. To improve the performance of the original CycleGAN [141] for this task a variety of adjustments were made, with the most significant being the introduction of a mid-cycle activation map loss.

My model contains two mapping functions $G_{Experimental}$ (G_{exp}): Simulated → Experimental and $G_{Simulated}$ (G_{sim}): Experimental → Simulated and associated adversarial discriminators $D_{Experimental}$ (D_{exp}) and $D_{Simulated}$ (D_{sim}). $D_{Experimental}$ encourages $G_{Experimental}$ to translate experimental images into outputs indistinguishable from real experimental images, and vice versa for $D_{Simulated}$ and $G_{Simulated}$. Both cycles include the cycle consistency loss that was introduced in the original paper Figure 27 (b, c). To further encourage accurate defect reproduction, I introduce a mid-cycle activation map loss for the simulated image cycle Figure 27 (b).

The mid-cycle activation map loss aimed to give the algorithm freedom to alter the noise profile whilst retaining constraint over the original defect response. The need for this was clear from the original implementation as the defect response can easily be washed out (Figure 25). To do this, the simulated input image was used to generate an activation map. This activation map was a normalised version of the original simulated input image to a range of 0 and 1. The simulated responses allowed for this unique implementation as the background responses were uniform. By normalising the activation map, the effect of background response was zeroed, and only inaccurate reconstructions of defect responses were punished, whilst maintaining even weak defect responses. Next, a scale factor was calculated to allow for adjustments of defect size. This was calculated by taking all non-zero values (defect response) from the activation map and dividing by the total image area. The L1 unreduced absolute error between the generated image and the simulated image was then calculated. The activation map was then applied to focus the loss to the defect response and minimise the loss from the noise. This new loss map was then divided by the scale factor previously calculated from the activation map. This means that the loss function is indiscriminate of defect size and does not punish larger defects more significantly than smaller defects. Finally, the mean was taken to get the reduced value, which was fed into the combined generator loss function given by (9). Figure 26 demonstrates this process with an example image.

$$L_{activmap}(G_{exp}) = E_{sim\sim P(sim)}\left[\frac{\|G_{exp} - sim\|_1 \times M_{activation\ map\ [0\to1]}}{K_{scale\ factor}}\right] \quad (9)$$

$$L_{total}(G_{exp}, G_{sim}, D_{exp}, D_{sim}) = L_{GAN}(G_{exp}) + \frac{2}{3}L_{GAN}(D_{exp}) + L_{GAN}(G_{sim}) + \frac{1}{3}L_{GAN}(D_{sim}) + \frac{\lambda}{3}\left(2L_{cyc}(G_{exp}) + L_{cyc}(G_{sim})\right) + 2\lambda L_{activmap}(G_{exp}) \quad (10)$$

Where L_{total} is the total loss, L_{GAN} and L_{cyc} represent the adversarial loss and cyclic loss given in the original paper [141], $E_{sim\sim P(sim)}$ represents the expectation over the batch of simulated samples, $M_{activation\ map\ [0\to1]}$ is the normalised activation map, λ is a coefficient to balance the relative importance of each loss function during training.

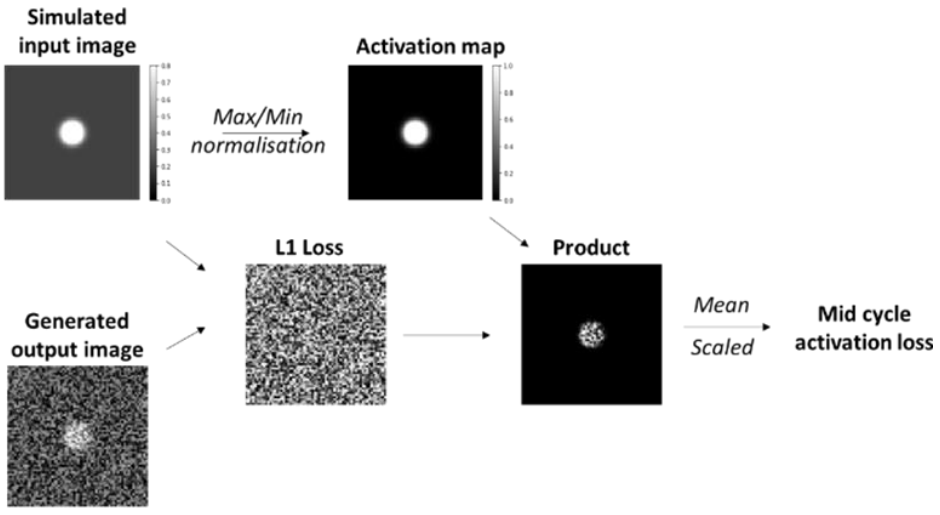
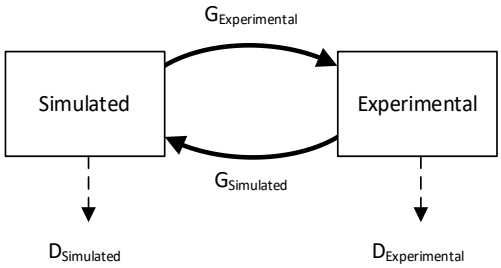
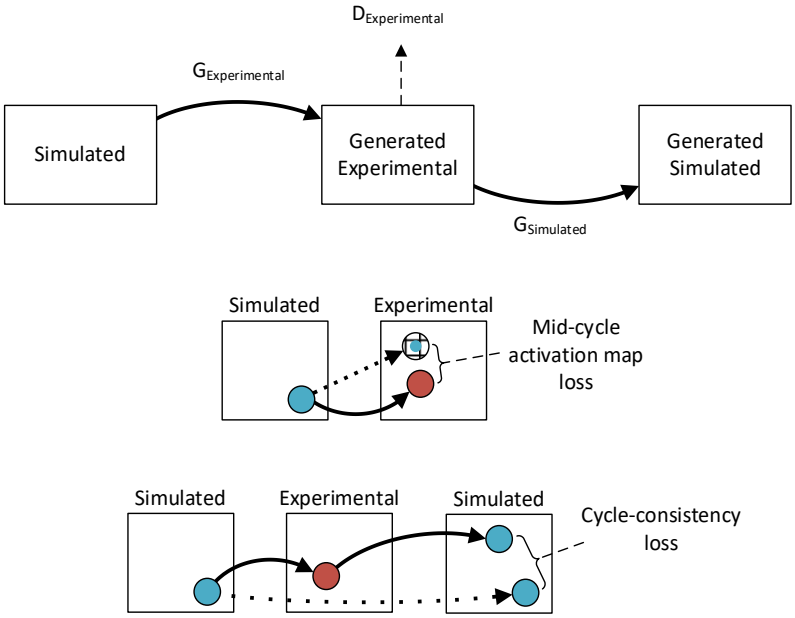


Figure 26: Diagram showing how an example mid-cycle activation map loss is generated.

a)



b)



c)

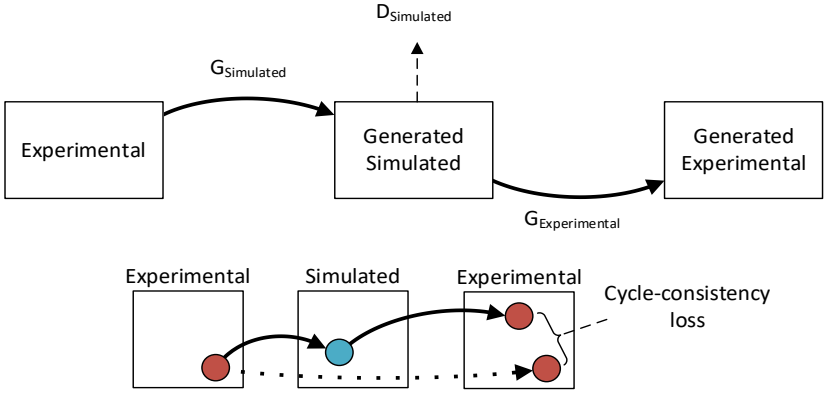


Figure 27: a) The model contains two mapping functions $G_{Experimental}$: Simulated → Experimental and $G_{Simulated}$: Experimental → Simulated, to transfer between the respective domains and the associated adversarial discriminators $D_{Experimental}$ and $D_{Simulated}$. b) When completing the full cycle from the simulated domain, the mid-cycle loss is added along with the cycle loss. c) When completing a cycle beginning in the experimental domain, the cycle loss is solely used as the mid-cycle loss is not calculable for the simulated domain.

The mid-cycle activation map loss is only applied in the direction going from simulated responses to generated experimental images, as it relies on the clean defect response of simulated images. This is demonstrated by Figure 27 (b, c). For better images in this task, the cycle loss was also adjusted to give twice the weighting of the simulated input cycle compared to the experimental cycle, whilst the discriminator loss for identification of experimental images was weighted twice as much as the discriminator for simulated images. This was done to further remove restrictions on noise generation and further encourage accurate defect response, whilst focusing on generation of experimental images over simulated images. The cycle loss coefficient (λ, equation (10)) was set to 100, with the mid cycle activation loss set to double the cycle loss. To further improve the results, the CycleGAN model used was adjusted from the original implementation [141] to perform better on the lower resolution 64x64 ultrasound images, by optimising the size of the first generator convolutional layers to 3x3 instead of 7x7, with 6 residual blocks used. The model was trained from scratch with a learning rate of 0.0002 which decayed linearly after 100 epochs to zero for the remaining training. For training, the GAN used the experimental defective train dataset of 334 images, and the simulated defective dataset of 154 images. The GAN model was trained over 2300 epochs using a batch size of 128 using an NVIDIA GeForce RTX 3090 and took less than 8 hours to train. All other parameters were unmodified from the original paper [141]. The GAN model was created using the Pytorch framework.

Once trained, the learnt mapping from the GAN was used to convert the original 154 simulated images to a new synthetic dataset of defective images. The synthetic dataset produced high quality ultrasonic amplitude images which are visually comparable to

experimentally obtained images, examples of images generated from their corresponding simulated input are shown in Figure 28.

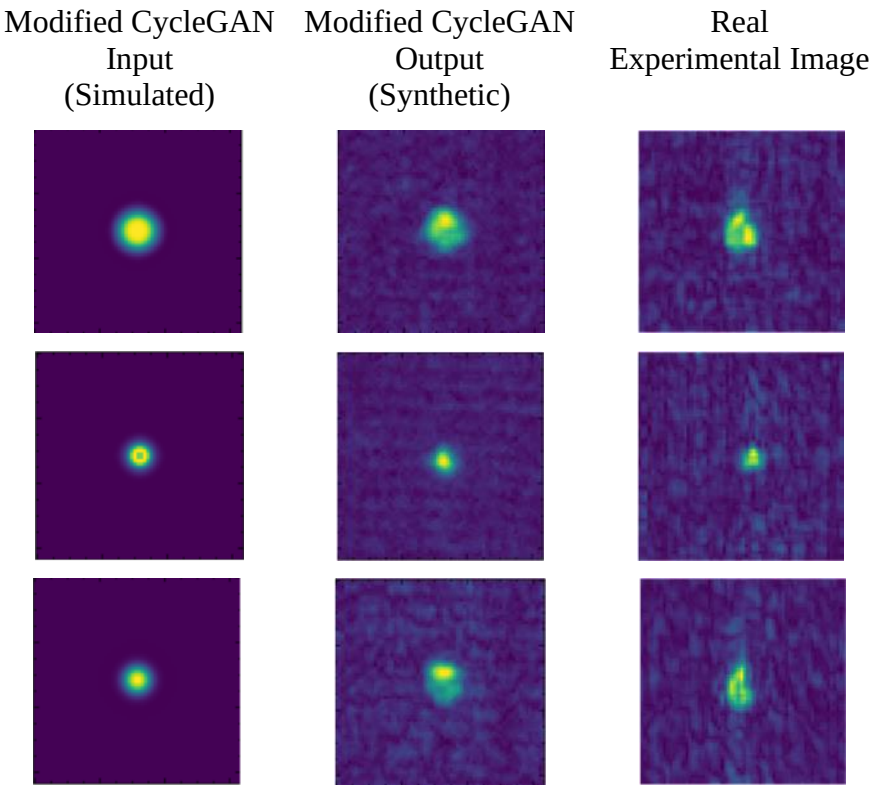


Figure 28: Example of synthetic generated images from their corresponding simulated defect input, along with real experimental images for comparison.

3.6.1.2 Classification Results

Training the CNN with GAN generated synthetic dataset and an equal number of clean images sampled from the clean train set, had a significant increase in classification performance compared to unprocessed simulated data when tested on the experimental clean and defective test datasets of 298 total images. After 100 training iterations, the model gave an average accuracy (± σ) of 87.0 ± 11.5%, (up from 62.8%) with average F1, precision and recall scores of 0.837 ± 0.204, 0.926 ± 0.231 and 0.775 ± 0.200 respectively. The average confusion matrix for the model is given in Table 9.

Table 9: Average confusion matrix across 100 training iterations for a CNN trained on GAN generated synthetic data.

		Predicted	
		Defect	**No defect**
True	**Defect**	144.6	5.4
	No defect	33.3	114.8

3.6.2 Approach 2: Experimental C-Scan Noise Superposition

Out of the 334 clean experimental C-scan images from the clean train dataset, 154 were randomly sampled to match the size of the simulated dataset. The simulated defect images were then combined with the real noise images by summation at an individual pixel level. To not exceed the normalised upper value limit of 1, if a pixel value exceeded 1 due to the addition of noise, it was clipped to remain within the limit. This was done instead of re-normalising the dataset as this would have reduced the noise distribution from the experimental data. From the new dataset, the images where the noise was greater than the signal were removed. This left 83 final images. An example of this is demonstrated in Figure 29.

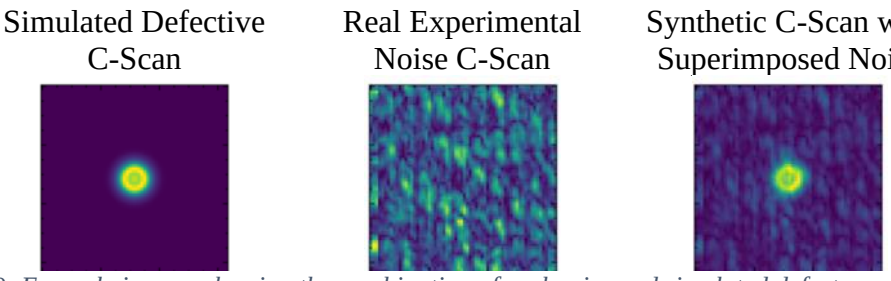


Figure 29: Example images showing the combination of real noise and simulated defect responses.

A considerable downside of the real noise approach is that it is not a fully simulated approach. This restricts its ability to scale as it requires an equal number of clean experimental images as simulated images. However, the experimental data required is from defect-free images which are more accessible and considerably easier to acquire

than real defect responses. The computational complexity of scaling this approach to a large number of images would be low. Therefore, if adequate clean images were available this technique could be used to produce a large dataset.

3.6.2.1 Classification Results

Training the CNN with the experimental noise synthetic dataset and an equal number of clean images sampled from the clean training set had a significant increase in classification performance when tested on the experimental clean and defective test datasets compared to the simulated data but was unable to match the results from the GAN generated dataset. After 100 training iterations, the model gave an average accuracy (± σ) of 77.4 ± 7.8%, with average F1, precision and recall scores of 0.688 ± 0.179, 0.950 ± 0.218 and 0.545 ± 0.158 respectively. The average confusion matrix for the model is given in Table 10.

Table 10: Average confusion matrix across 100 training iterations for a CNN trained on real noise data.

		Predicted	
		Defect	**No defect**
True	**Defect**	150.0	0.0
	No defect	67.3	80.7

3.6.3 Approach 3: Simulated C-Scan Noise

To reduce the experimental demand of the real noise superposition approach requiring a unique experimental image for each simulation, a study was conducted to understand if it was possible to fully simulate the experimental noise profile. To do this, the noise distribution from the clean experimental C-scan images of the defect free sample were analysed by plotting a histogram. It can be seen from Figure 30 that this noise profile

is well aligned with an inverse gaussian distribution given by μ 0.410, loc -0.003 and scale of 0.066, the distribution was implemented using SciPy, with details of the distribution given by [148].

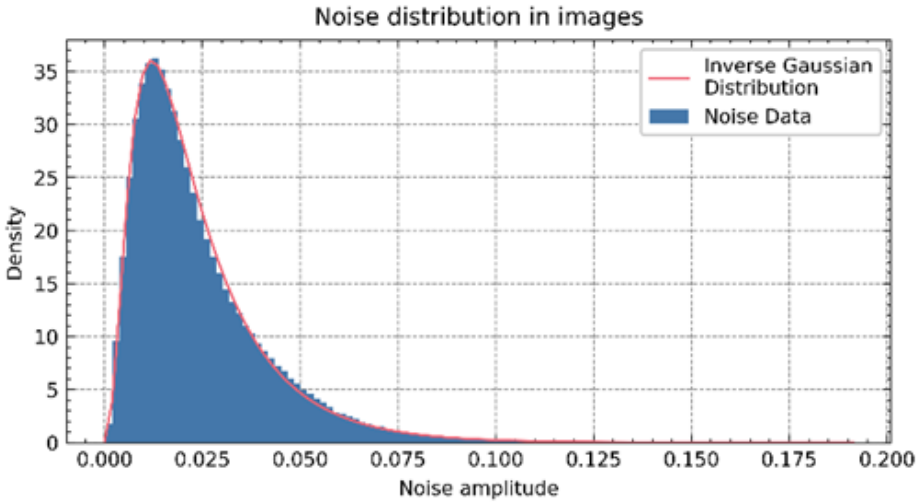


Figure 30: Density histogram showing the distribution of data from the clean sample.

The simulated defect images were then combined with a noise pattern which was randomly generated for each image from an inverse gaussian distribution, Figure 30, with the previously determined parameters. The images were combined by summation at an individual pixel level. As per the real noise method, to not exceed the normalised upper value limit of 1, if a pixel value exceeded 1 it was clipped to remain within the limit. From the new synthetic dataset, the images where the noise was greater than the signal were removed, and we were left with 80 C-scan final images. An example of this is demonstrated in image Figure 31.

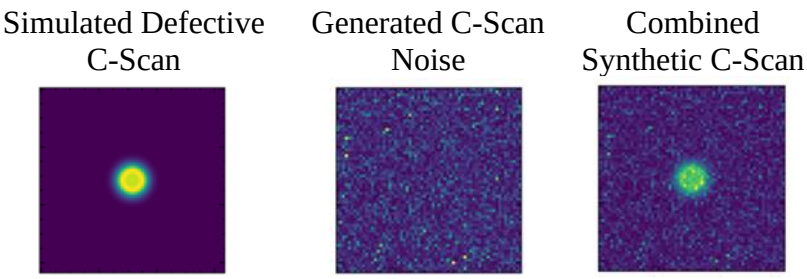


Figure 31: Example images showing the combination of C-scan simulated noise and simulated defect responses.

The implementation of C-scan noise at scale would be considerably easier than the real noise approach. This is as fully simulating the noise profile from an appropriate experimental distribution requires little additional experimental data acquisition after a suitable population has been sampled. Furthermore, the computational complexity of this implementation is as efficient as the real noise approach and could scale well to produce a large dataset. Whilst it benefits from simplicity, the approach could be extended to account for local correlations between pixels, to more accurately simulate local relationships resulting from the composite ply structure.

3.6.3.1 Classification results

Training the CNN with the C-scan noise synthetic dataset and an equal number of clean images sampled from the clean training set produced poorer results than the superimposed real noise dataset but still improved significantly in classification performance over the simulated dataset when tested on the experimental clean and defective test datasets. After 100 training iterations, the model gave an average accuracy ($\pm \sigma$) of 74.3 ± 8.1%, with average F1, precision and recall scores of 0.629 ± 0.195, 0.930 ± 0.255 and 0.482 ± 0.164 respectively. The average confusion matrix for the model is given in Table 11.

Table 11: Average confusion matrix across 100 training iterations for CNN trained on simulated C-scan noise data.

		Predicted	
		Defect	**No defect**
True	**Defect**	150.0	0.0
	No defect	76.7	71.3

3.6.4 Approach 4: Simulated Ultrasonic A-Scan Noise

An approach of fully generating a simulated noise profile at an A-scan level was also investigated which is better aligned to how noise occurs from the physical response of ultrasonic signals. For each individual time trace signal, the complete noise profile is composed of both structured noise and random noise. Structured noise consists of physically accurate responses, just not from a known feature. These are likely due to the interaction of different composite plies and the component geometry with the propagated ultrasonic waves. Whereas random noise is independent of the samples structure and could be due to random electrical noise for example.

It was assumed that for a given B-scan, the structural noise profile will remain constant, as for a given B-scan the ultrasonic wave and ply layer interactions and therefore backscattering noise should be similar. Therefore, at a B-scan level, it is possible to remove most of the random noise by mean averaging the individual A-scans together at each sample interval leaving the structural noise component. For each A-scan in each B-scan, it is then possible to work out the random noise component from the differences between each A-scan and the structural noise component on a per sample basis. These combined differences can be plotted on a histogram to represent the random noise population of a B-scan. This process was completed for each individual B-scan. The random noise profiles were combined to give a greater number

of samples for the distribution. From Figure 32, it can be seen that this distribution is approximated by a normal distribution with 0.000 mean and a standard deviation of 0.013.

To learn the variation of the structural noise components across B-scans, the average B-scan structural noise was first calculated by averaging each individual B-scan noise profile on a mean sample basis. The difference between the mean and each individual B-scan structural noise profile was calculated on a per sample basis and again plotted on a histogram (Figure 33). This can be approximated by a normal distribution with mean 0.000 and standard deviation 0.003.

To generate a new noise pattern for a B-scan, a new structural noise pattern was generated by taking the overall mean structural noise pattern and adding variation based on the normal distribution previously calculated. To make this signal more representative of the Hilbert transformed A-scan data, a Savitzky–Golay filter [149] was applied to smooth the data (Figure 34). Afterwards, a random noise profile was added to the generated A-scan baseline signal, following the previously determined normal distribution for each A-scan in Figure 32. Figure 35 helps to illustrate this process at A-scan and B-scan levels. The simulated responses were then combined with the generated combined noise profiles using a per sample summation. As per previous methods, to not exceed the normalised upper amplitude value limit of 1, pixel values exceeding 1 were clipped to remain within the limit. From the new dataset, the images where the noise was greater than the signal was removed resulting in 126 C-scan final images. An example of the final images is demonstrated in Figure 36.

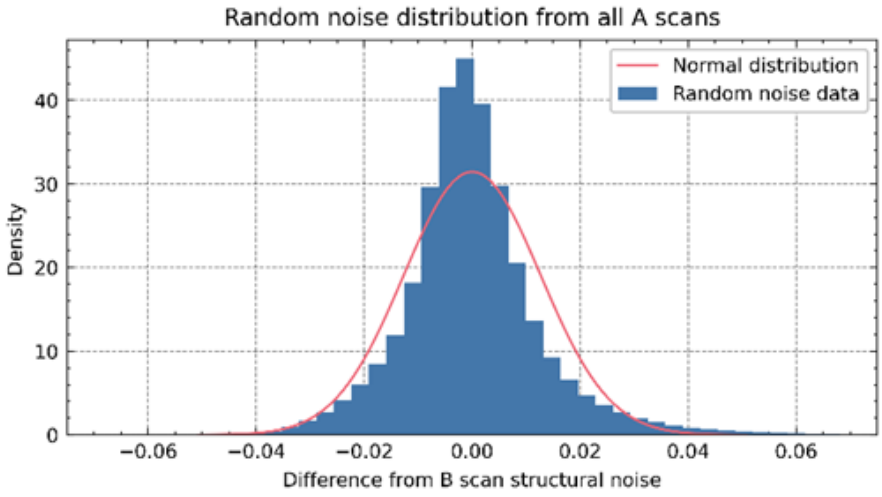


Figure 32: Density histogram showing the random noise distribution from the total A-scans.

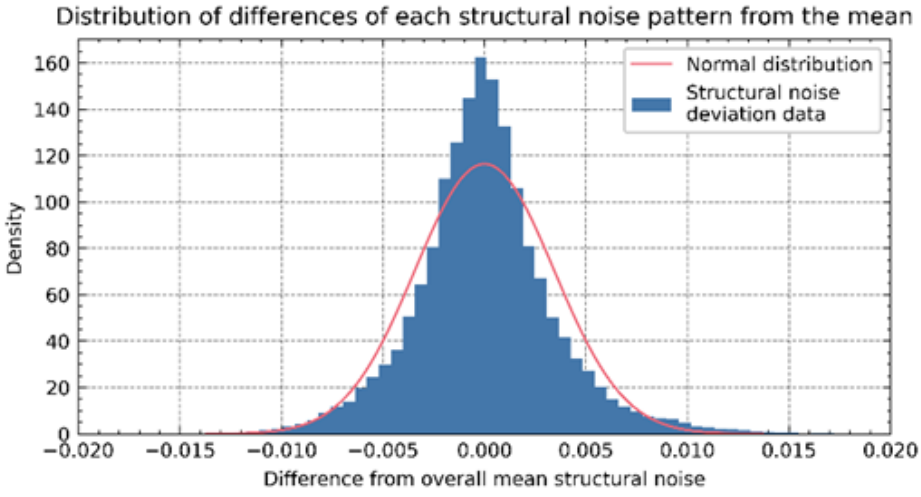


Figure 33: Density histogram showing the distribution of deviation for structural noise from the mean structural noise pattern.

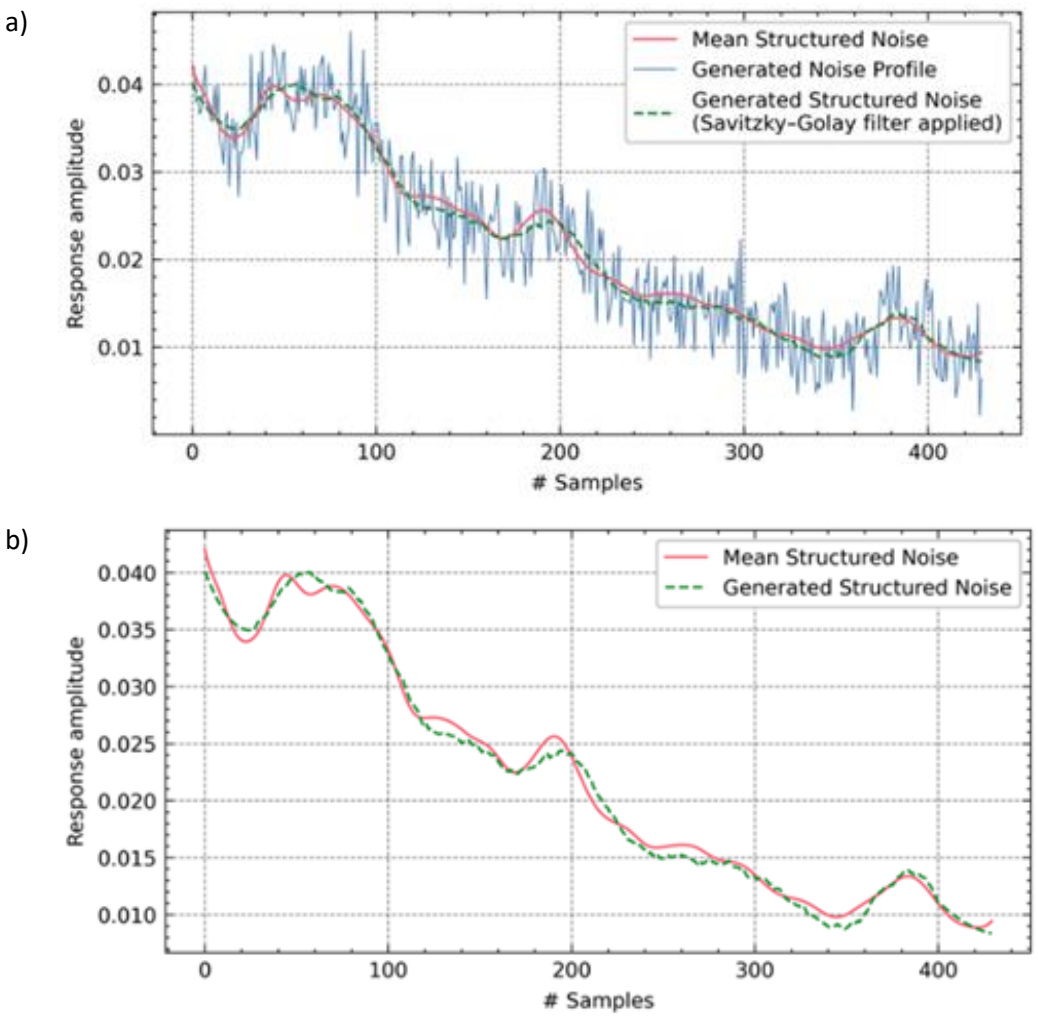


Figure 34: a) An example of how a structural noise profile is generated from the mean. b) A cleaner example of the final generated noise profile.

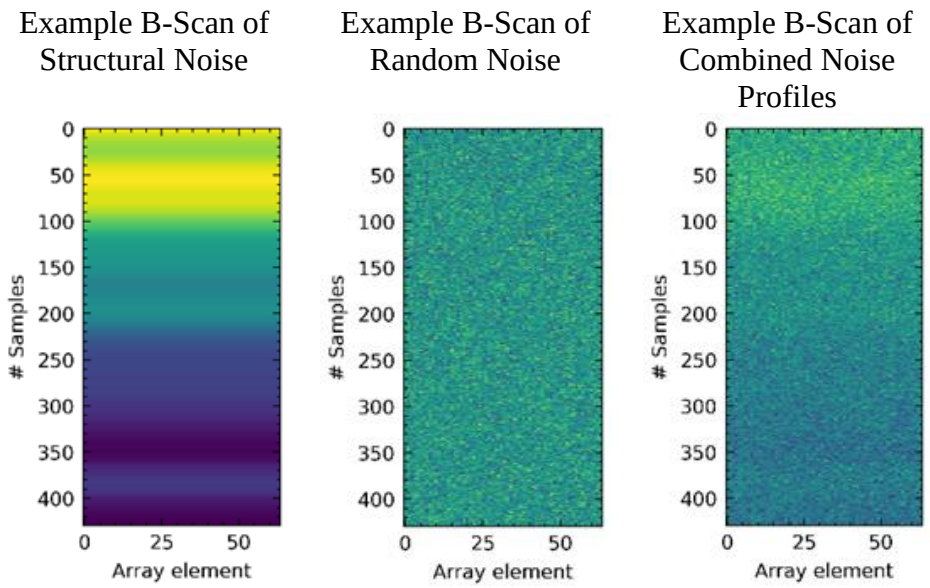


Figure 35: An example of how structural and random noise profiles are combined at a B-scan level.

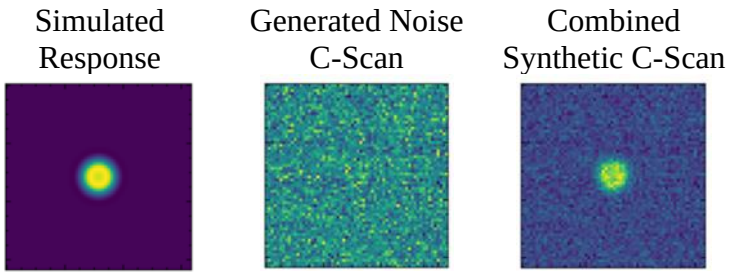


Figure 36: Example images showing the combination of A-scan simulated noise and simulated defect responses.

Whilst implementing the A-scan noise profile does require experimental analysis and characterisation, the application to simulated data is a fully simulated approach. In addition, by adding noise at an A-scan level, it allows for the potential of three-dimensional volumetric analysis, or analysis of B-scan images, which is not possible with any of the other methods. However, it requires a greater level of analysis compared to the C-scan level noise method before implementation. Furthermore, as the generation of the noise pattern is required on a per B-scan level, an additional computational step is required to cover the number of B-scans. This is therefore less computationally efficient than both the real noise and C-scan noise implementation.

3.6.4.1 *Classification results*

Training the CNN with the A-scan noise synthetic dataset and an equal number of clean images sampled from the clean training set, gave the second-best classification performance when tested on the experimental test datasets after the GAN generated dataset. After 100 training iterations, the model gave an average accuracy (± σ) of 80.0 ± 6.2%, with average F1, precision and recall scores of 0.738 ± 0.141, 0.970 ± 0.171 and 0.598 ± 0.124 respectively. The average confusion matrix for the model is given in Table 12.

Table 12: Average confusion matrix across 100 training iterations for a CNN trained on simulated A-scan noise data.

		Predicted	
		Defect	**No defect**
True	**Defect**	150.0	0.0
	No defect	59.5	88.5

3.7 Discussion

3.7.1 Comparison of Classification Results

Figure 37 shows examples of C-scan images produced by the different synthetic data generation methods. The classification results are summarised in Figure 38 and Table 13, which show the mean (μ) and standard deviation (σ) accuracy and F1 scores, and full evaluation metrics respectively for each dataset investigated.

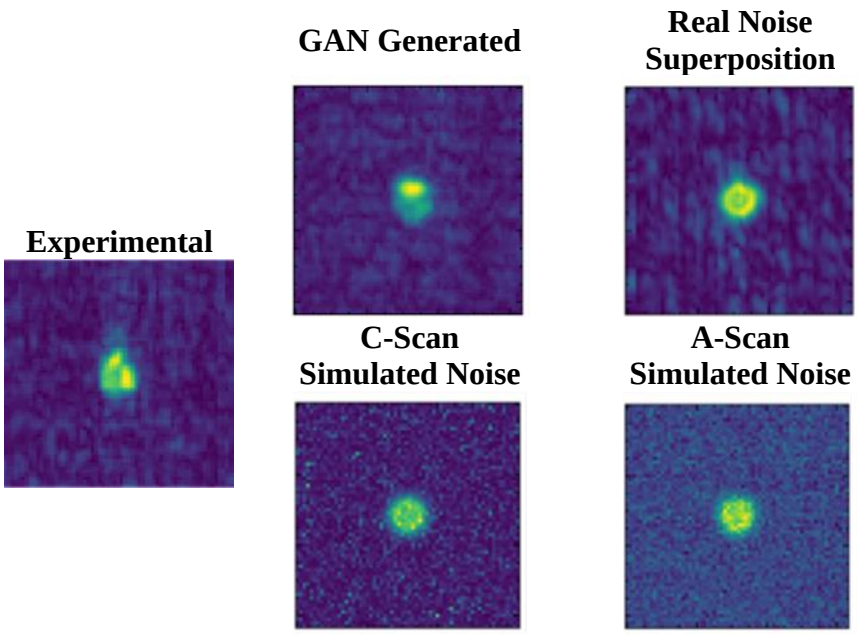


Figure 37: Comparison of different real and synthetically generated C-scan image examples.

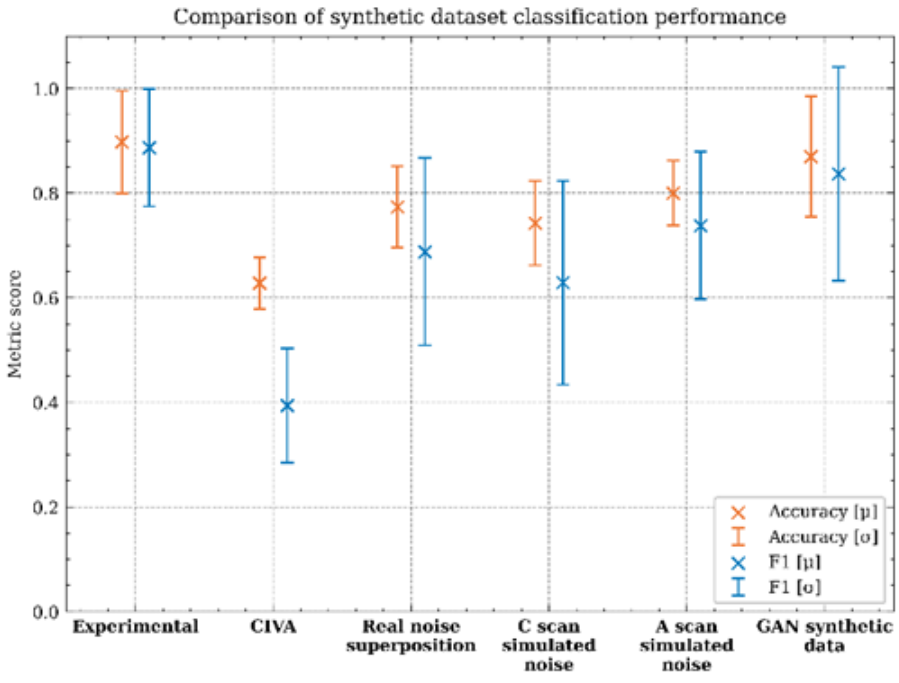


Figure 38: Comparison of classification results for each dataset.

Table 13: Summary of classification results for each dataset.

Training dataset	Evaluation metric			
	Accuracy	**F1**	**Precision**	**Recall**
Experimental	89.8%	0.887	0.974	0.826
CIVA	62.8%	0.394	1.00	0.252
Modified CycleGAN	87.0%	0.837	0.925	0.775
Real Noise Superposition	77.4%	0.688	0.950	0.545
C-Scan Simulated Noise	74.3%	0.629	0.930	0.482
A-Scan Simulated Noise	80.0%	0.738	0.970	0.598

Simulated UT data of defect responses in composites lacks the complexity of experimental noise. In this work, it was demonstrated that when CNN classifiers are trained on purely simulated data and tested on real experimental data a significant adverse impact on classification performance is observed, with an average F1 score of 0.39. However, four novel strategies were proposed and explored in this research for creating synthetic composite UT datasets to reduce this effect with the aim to better simulate real experimental data. According to the results of this study, all four methods showed significant increases in classification performance compared to the original simulated dataset. Among these, the modified CycleGAN generated synthetic dataset produced significantly better classification results than the other methods, with an average F1 score of 0.84. This neared the classifier trained on a subset of the experimental dataset, but due to the reduction in available experimental training and test data due to the train/test split this should not be considered a direct comparison. For direct comparison an additional experimental dataset would have been required for training and testing on the complete test set.

Superimposed experimental noise, simulated C-scan noise, and simulated A-scan noise produced similar mean accuracy results, but the simulated A-scan noise synthetic dataset produced the best average F1 score of the three, with 0.74. It is interesting that the simulated A-scan noise dataset outperformed the real noise synthetic dataset. This may be due to the fact the real noise obscures the defect response features too much. Alongside the ability to accurately simulate noise response, a further reason for improved classification results for GAN and A-scan synthetic datasets may be their ability to account for depth wise signal attenuation and adjust the noise levels with respect to depth and signal response. This produces more appropriate noise levels for deeper and weaker defect responses and allows for the preservation of many more simulated responses. Unlike simulated C-scan and real noise approaches which are defect depth agnostic and therefore result in the rejection of more images due to the concealment of low-level responses with noise profiles that are not depth matched. These methods could be extended to include a finer depth wise noise implementation, but this is outside the scope of this work and is left for future investigation.

These results demonstrate that in scenarios where noisy experimental environments can cause real data to vary greatly from simulated data, synthetic methodologies for noising data provide an opportunity for generating more effective training data. This is particularly beneficial as we retain the accuracy and fully labelled nature of physics-based simulations, which allow us to fully control the simulation of different defect class types and the variability within them.

When considering the broader aim of generating large synthetic datasets that could be used to create a database of realistic training examples, it is important to consider the

ease and robustness of synthetic data generation. Training of the CycleGAN is a delicate process and whilst it has been able to produce realistic images for FBHs, it may struggle to generalise to other defects without significantly broader examples of defects in training. This would largely defeat the point of the synthetic data generation in this instance. Furthermore, the training of an effective GAN model is still extremely challenging and the process of hyperparameter selection is not robust. It is therefore favourable to consider an approach that is robust to different defect types and can be scaled. For scalability, a fully simulated method is preferable over a method which still requires significant collection of experimental data. Therefore, the real noise approach is superseded by both the A-scan and C-scan synthetic approaches. The C-scan noise approach is slightly easier to implement than the A-scan as it requires less experimental data analysis and can be done at the C-scan image level instead of the A-scan level. However, the A-scan noise approach allows for noising of the full volumetric data, which could provide opportunities in three-dimensional data analysis. Further work could be done to explore the distribution of C-scan noise at different depths to enable maintenance of a larger number of simulated responses in a simpler way than the complex A-scan noise simulation method. This could potentially combine some of the benefits of both the A-scan and C-scan noise approaches but would remove the opportunity for volumetric data analysis if done at an image level. In certain scenarios, gathering clean experimental data may not be a limiting factor and in this case, it could be beneficial to expand the real noise superposition method to align the images depth wise between the experimental and simulated domains. This would help to better account for localised structural noise, and likely give improved classification performance similar to the A-scan noise approach.

3.7.2 Model Interpretability with Guided Grad-CAM

A key barrier to the uptake of Machine Learning in NDT is a lack of model interpretability [14] and the use of synthetic data has the potential to further mystify this process. To help tackle this issue, Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) was implemented for a randomly selected model trained from each dataset and evaluated on experimental data. Guided Grad-CAM is a technique for producing 'visual explanations' of CNNs with the goal of making them more transparent and explainable [150]. Guided Grad-CAM gives a visual indication of what inputs are used for positive class prediction. Whilst it does not give any information about how or why the inputs are used for the prediction within the model it has been shown to help users place greater trust in a model. The method combines Guided backpropagation and Class Activation Maps (CAM) to create visualisations which indicate relevant image regions for class-discriminative predictions. Guided Grad-CAM is not a complete solution for model interpretability; however, the goal is to visually compare if the models trained on synthetic data are using similar inputs for prediction compared to models trained on experimental data, in the hope that this provides trust in the viability of using synthetic datasets. Figure 39 shows the defective experimental test image, and both the associated Guided Grad-CAM image which gives a visual indication of significant regions contributing to defective predictions, and a mixed image which combines the Guided Grad-CAM and the input image with a respective weighting of 1.5.

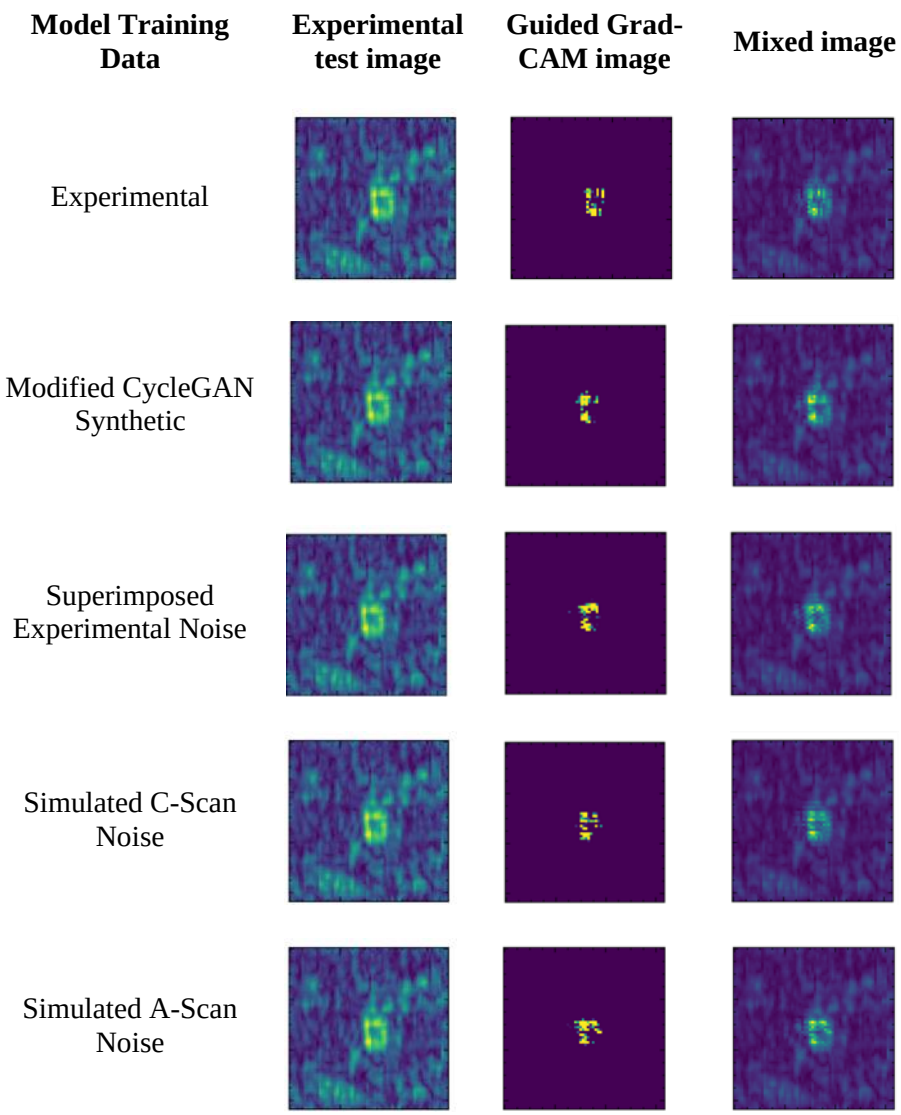


Figure 39: Example of Grad-CAM visualisation of models trained on different datasets.

It has been identified in literature that model interpretability is a key limiting factor in the uptake of DL in NDT. Guided Grad-CAM was implemented to try and minimise the obscurity that using synthetic data could produce. Whilst model interpretability is a complex field of research and interpretability is challenging to quantify, it is hoped that the Guided Grad-CAM results at least indicate that models trained on synthetic data are learning similar features compared to models trained on purely experimental data.

The Grad-CAM images consistently and accurately highlight defect pixels in the context of defect detection. The substantial similarity between Grad-CAM results from models trained on synthetic data and those trained on experimental data suggests that both types of models focus on the same relevant features. This consistency across different training data sources enhances confidence in the models' predictions and supports the use of synthetic data as a viable alternative to experimental data for training.

By providing clear visual evidence of accurate defect region identification, Grad-CAM helps to address concerns about the reliability and validity of using synthetic data. This assurance increases confidence in deploying DL models trained on synthetic data for practical NDT applications.

3.8 Conclusion

Deep learning provides an attractive solution for helping to automate the interpretation of ultrasonic testing NDT data results in quality assurance processes. A barrier to implementation is that DL approaches typically demand large quantities of representative training data to allow accurate and reliable predications to be established. However, since modern manufacturing processes strive to reduce the incidence of defect formation, there is a paucity of real-world defect data available for ML training. Despite this work focusing on FBHs, by employing physics-based simulations of ultrasonic response to defects, it is possible to generate large sets of defect data, with variability in defect types, sizes, and orientations. A drawback in such simulation is in replicating the same noise distributions encountered in experimental measurements, and this is challenging without increasing model complexity to the

point of computational intractability. In this study, 4 techniques to map the noise distribution of experimental data onto our simulated data were presented to improve the performance of subsequent ML based classification of defects. A generative network was used to learn the mapping between simulated and experimental images, this resulted in a mean F1 score of 0.843. A method of combining clean experimental images with simulated images was introduced which resulted in a mean F1 score of 0.688. To remove the requirement for clean experimental images, two methods of fully generating synthetic noise profiles, C-scan and A-scan noise, were presented; the latter being based on a closer physical representation of how noise is produced experimentally. These methods produced mean F1 classification results on experimental data of 0.629 and 0.738, respectively. Whilst each method produced a significant improvement in classification over the purely simulated data, with a modified loss function to encourage accurate defect response, CycleGAN showed the greatest improvement in classification performance, allowing us to maintain the utility of simulating data from physics-based models and convert them to more experimentally realistic synthetic datasets. However, it was identified that other synthetic data generation methods may be more appropriate for generating large datasets, such as A-scan noise due to their greater robustness.

Model interpretability is a significant challenge for the uptake in use of Deep Learning in UT, with the use of synthetic data likely to further add ambiguity. To help minimise this, Guided Grad-CAM was implemented which visually indicated that models trained on synthetic data were learning similar features to models trained on experimental data for classification. This aids in providing confidence that the methods of generating synthetic data are appropriate for training experimental classifiers.

Whilst classification results for individual synthetic datasets had room for improvement, this work demonstrates that the synthetic data generation methods were able to successfully transfer the simulation domain closer to the experimental domain. This demonstrates a viable approach to training DL models when experimental data is unavailable, as with many NDT applications.

Future work will look to maximise the classification accuracy of specific models. This could be done by combining this work with additional domain adaption techniques, which have shown promise in previous literature [99]. Further investigation will also be conducted to optimise individual model classification accuracy by performing HPO directly using synthetic datasets. This would demonstrate the effects of performing HPO on a model trained on a synthetic dataset and whether this improves its classification in the experimental domain. This would also eliminate the need for experimental data entirely when training a DL classifier as both the parameter optimisation and training could be conducted in the fully synthetic domain. This would require only a small amount of experimental data for testing. Additionally, the next steps in this work will look to see if the style transfer can be extended across the full range of defect types and tested on naturally occurring experimental defects. It would also be beneficial to identify if it is possible to detect more challenging defects such as superficial defects using similar methods or if the UT approach would require modification. If successful, large, fully annotated, synthetic datasets could be efficiently produced, opening the potential for further use of DL in NDT.

4 Volumetric Detection

4.1 Introduction

Synthetic datasets are widely used in ML to augment small training datasets [136] and they have been successfully implemented for UT of composites with encouraging results for 2-dimensional (2D) classification of C-scan images in the previous chapter. Part of this work builds upon the work in the previous chapter to extend one of the synthetic data generation methods to make it applicable for full 3D volumes. The synthetic datasets are based on simulations from semi-analytical physics based software that has been shown to produce experimentally accurate defect responses [133], [134]. This software offers a less computationally expensive alternative to FEA, allowing for the simulation of composite responses based on bulk material properties [135].

When ML is used to interpret UT NDE data in literature, it is typically applied to interpret A-scan time traces or 2D images constructed from A-scans [97], [98], [101], [104], [106], [131], [132]. Compared to B-scans, A-scans lack all spatial information and nowadays, they are rarely used alone to characterise defects by human operators since the introduction of phased arrays. C-scans preserve detailed spatial information, however constructing the 2D image from the volumetric data necessitates the compression of temporal information. Whilst C-scans excel in capturing intricate spatial details, their need for temporal compression results in minimal representation of through-depth features. Compression of A-Scans to C-Scans often removes useful features such as the backwall response, which can be important when detecting defects with a low reflective index such as porosity [151]. Furthermore, to produce C-scan

images appropriate gating must be applied to remove the front wall surface response. This can be challenging when trying to detect near-surface defects. In the aerospace industry, operators typically start with a C-scan to gain a complete picture of defect responses and then move to analysis of B-scans for further information about the nature of the responses [152], [153]. Whilst current ML approaches in literature make use of data in formats that are easily interpreted by humans (images or time-traces), ML algorithms are not limited to image-level analysis and have proved very capable at interpreting 3D volumetric data [154], [155]. By implementing algorithms capable of volumetric interpretation, we retain all spatial and depth information, this gives the algorithms more relevant features to learn from and removes the need for image pre-processing and gating.

CNNs have been used effectively for decades in a wide variety of image and volumetric analysis tasks with models such as ResNet typically having tens of millions of parameters [156], and are still widely used as backbones or standalone architectures [74]. However, these networks are typically applied to data of similar dimensions, or data which has been scaled to give even dimensionality of each axis. UT data has extreme aspect ratios due to the difference in requirements of sample rate in the spatial and time dimensions. The upper limits of the temporal and spatial resolution are determined by the sample thickness and by factors such as the number of elements in the ultrasonic probe or the scan length, respectively. Compressing the data in the time dimension to match the spatial dimension, normally dictated by the sub-aperture pitch and the scan acquisition rate, would result in a substantial loss of depth information. Alternatively, the spatial dimensions could be upscaled to match the number of samples in the time dimension, but this is highly inefficient, creating data instances

that would require large amounts of memory, and would make training intractable. Therefore, retaining the original dimensionality and aspect ratio of the UT data is highly preferable. Using CNNs to interpret images with high dimensionality is not new and the use of rectangular kernels instead of square kernels in CNNs has given positive results for classification of speech signals, which have high aspect ratios when represented as spectrogram images [157]. The work in this chapter makes use of a similar approach for volumetric data.

Network architecture design is a key component of effectively leveraging machine learning techniques. Traditionally, network design heuristics and 'rules of thumb' would be used, in tandem with domain expert knowledge to construct a specific architecture. Automatic architecture design or Neural Architecture Search (NAS) is a development on this approach where a practitioner can leverage compute to aid the process of architecture selection. This process, which can be considered a subset of hyperparameter optimisation, generally involves an iterative process of selecting, training, and evaluating architectures. In its simplest form, a 'Random Search' involves repeating the above process until some threshold or limit in terms of performance or computation budget is reached. More complex approaches to NAS often focus on efficient model evaluations, making use of proxy evaluation methods [158], [159] or efficient sampling algorithms [160], [161] attempting to make the largest improvement with each evaluation.

This chapter presents a comparative analysis of the performance achieved from three separates architectures for defect detection in volumetric ultrasonic data. The first, VoxNet [162], is prevalent in the literature for volumetric classification problems, the

second architecture presents modifications to VoxNet for this task using a traditional network design approach, and finally a discovered architecture from NAS.

VoxNet is a 3D CNN initially proposed for classification of LiDAR, RGBD and CAD data. It has since been used as a backbone for different volumetric classification tasks [163]. Additionally, notable contributions of this study to knowledge in the field encompass the introduction of domain-specific augmentations, which exert a substantial impact on the classification performance. Furthermore, synthetic data generation techniques are leveraged from prior 2D work to generate 3D UT datasets from semi-analytical simulations, effectively addressing one of the prominent challenges encountered in the application of deep learning for NDE: the scarcity of effective training data.

This work presents a novel DL architecture designed to process volumetric UT data. In contrast to prior methods relying on time-series data or 2D image-based approaches, which diminish spatial or temporal features, whilst often requiring additional processing. The main contributions of this work are:

- Interpretation of volumetric UT data, instead of images or time signals. This reduces preprocessing requirements and allows the model to learn from greater features.
- Introduction of two domain specific methods for data augmentation, helping with the domain transfer from synthetic to experimental data.
- Discovery of a novel 3D CNN architecture through NAS.

4.1.1 Procedure

In this work, the automated data interpretation is simplified by inspecting the complete volumetric data, eliminating image processing steps like gating to remove front and back wall responses, while preserving all spatial and temporal information. Whilst the models are trained on synthetic data, they are tested using experimentally collected UT data from samples with manufactured defects that aim to mimic delaminations. Manufactured defects are commonly used in literature to act as test cases and qualify NDE techniques and operators where naturally occurring defects are not always available [32], [131], [132]. An overview of the simulation and deep learning pipeline is presented in Figure 40. Figure 40 also shows how NAS can fit into this process, with Figure 48 providing a more detailed overview of the NAS pipeline.

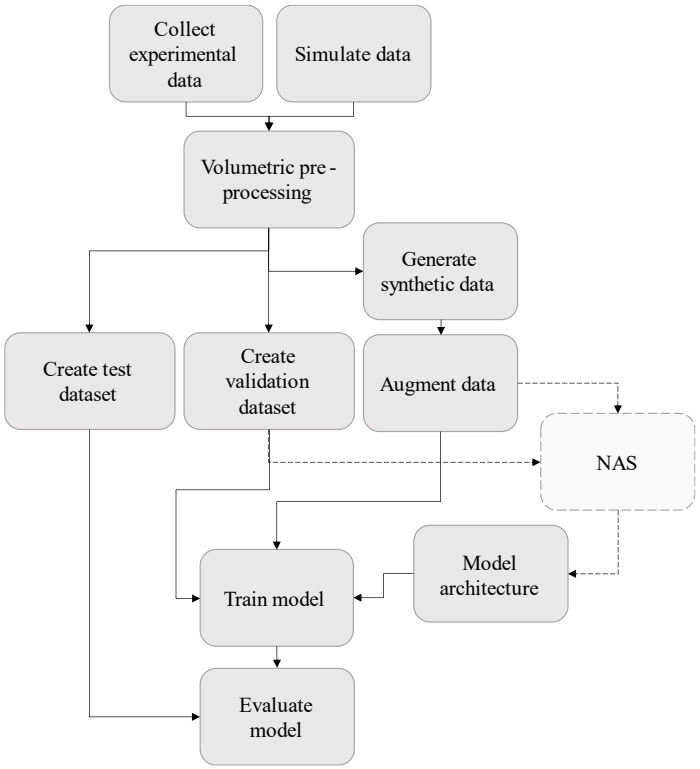


Figure 40: Overview of the pipeline for automated volumetric UT classification.

4.2 Data

4.2.1 CIVA Simulations

Due to the lack of available experimental training data, a simulated dataset was constructed for training. This was done using CIVA, a semi-analytical physics-based commercial NDE simulation software [142]. CIVA has the ability to model wave propagation and interactions with defects. It has been validated, showing good agreement to experimental results for different UT scenarios [133], [134]. Additionally, the software is computationally efficient when compared to other alternatives such as Finite Element Analysis (FEA). The full control of the simulated domain enabled the modelling of similar defects and material properties to the

experimental domain. However, the use of semi-analytical software instead of FEA had limitations in that the software was unable to model responses from ply interactions and lacks noise seen in experimental data. As a result, differences existed between the simulations and measured experimental responses, leading to the use of the synthetic data generation steps discussed in Section 4.2.2 to reduce the differences between simulated and experimental domains.

To set up the simulation, the individual layers of composite were constructed and used to generate equivalent homogeneous material properties of the experimental CFRP samples. A single ply layer was constructed and alternated repeatedly with 8 layers at orientations of 0, 45, -45, and 90 degrees to match the experimental sample as closely as possible (as given in section 2.2). The resulting multilayer structure was homogenised to be consistent with a homogeneous medium having mechanical properties equivalent to those of the multi-ply composite, with the fibre density set to 50% best match the experimental sample's density of 1440 kg/m^3. To simulate the waveform, a sinusoidal wave of 5 MHz was employed, accompanied by a Hanning filter that provided a bandwidth of 66% at 12 dBs.

For running multiple, sequential simulations, a parametric study was set up, using the composite bulk properties previously calculated and varying the diameter and depth of defects. FBH defects were simulated with diameters from 3.0 mm to 15.0 mm, increasing every 0.5 mm, with varying depths from 1.5 mm to 7.0 mm from the surface, in increments of 1.5 mm. A defect-free simulation was also run to provide the basis for defect-free synthetic data. Both the front and back wall surface reflections were included in the model. The full simulations took less than 15 hours on a desktop

computer with a 24-Core 3.79 GHz CPU and 128 GB of memory. An example of a simulated defect is shown in Figure 42 (a).

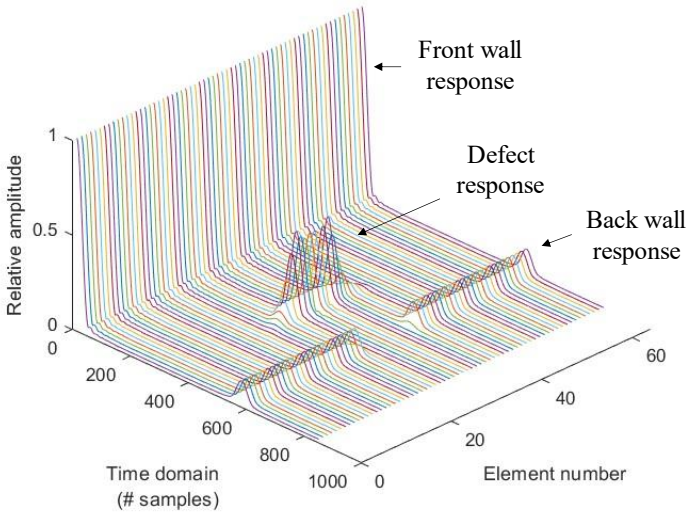
4.2.2 Synthetic Data Generation

Work conducted in chapter 3 has shown that semi-analytical simulated data alone is not representative enough of the experimental domain [164]. Therefore, there is a need for methods of translating the simulated domain closer to the experimental domain. Fully statistical methods of generating noise are advantageous as they can be re-sampled continuously to keep generating unique noise profiles which are in line with experimental data. In this work, we extend previous work in generating 2D synthetic images and propose a new approach for adding noise to complete volumetric UT data.

The previous study concluded [164] that A-scan level noise was the best fully generative statistical method for adding noise. Additionally, all the other approaches, except for the simulated A-scan noise, introduced noise at an image level, which is intractable for volumetric data. To adapt the methodology described in the previous chapter for the analysis of full volumetric data, unique noise profiles for each A-scan were generated and subsequently summed with the simulated responses past the front wall.

Figure 41 shows an example of the addition of noise on simulated data at an A-scan level and Figure 42 demonstrates this for a complete ultrasonic volume. The statistical noise distributions of the A-scans were calculated from a separate hold out sample with the same layup and thickness as the test samples. For further details on building up the noise profiles, please refer to the previous work [164].

a)



b)

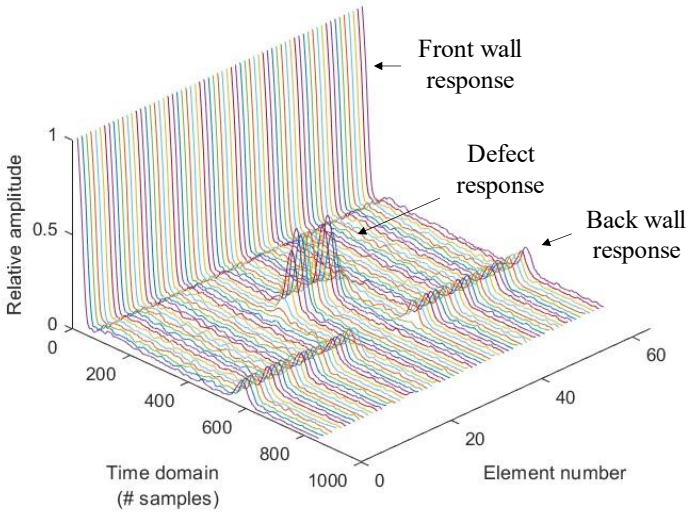


Figure 41: a) A frame of 64 simulated A-scans for a simulated defect response. b) the corresponding A-scans with synthetically added noise for the same defect response.

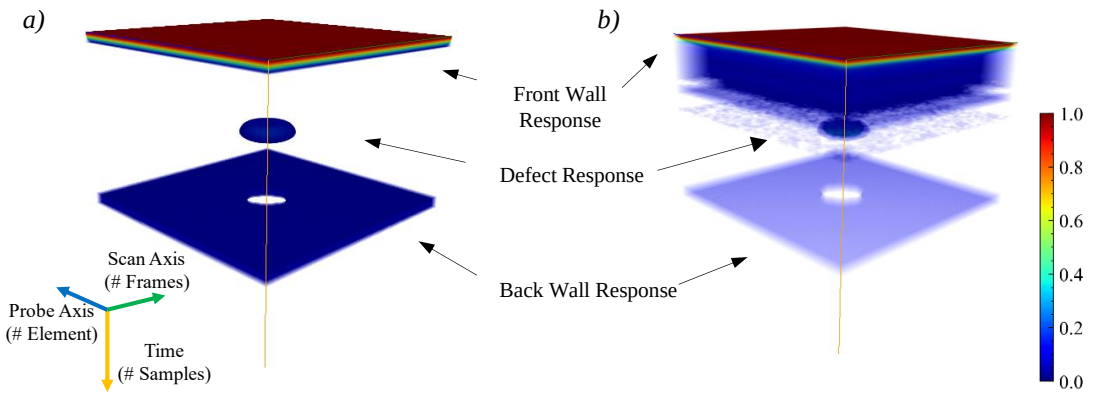


Figure 42: a) Complete ultrasonic volume of simulated A-scans for a defect response. b) the corresponding synthetically noised volume for the same defective response. Both figures have been thresholded to remove the lowest 10% of amplitudes to aid in visual clarity.

A summary of the datasets generated from the experimental and synthetic data is given in Table 14.

Table 14: Summary of the datasets produced.

Data source	Dataset	Number of datapoints
Simulated defect responses (300 Flat-Bottom Holes)	Synthetic defective train	300
Simulated defect free response	Synthetic defect free train	300
Experimental defect reference sample (15+25 Flat-Bottom Holes)	Defect test (70%)	25
	Defect validation (30%)	15
Experimental defect free reference sample	Defect free test (70%)	25
	Defect free validation (30%)	15

4.2.3 Augmentation

The generalisability of ML models is a critical aspect of their performance. One approach to improve generalisability is to augment the training data. Augmenting the training data makes the task more challenging by adding noise at the training stage, reducing the likelihood of overfitting, and often improves performance in the target domain. This is particularly important when the target (experimental) domain is different from the training (synthetic) domain.

As demonstrated in Figure 42 (a), there is little variation between simulated A-Scans. However, this is not the case for experimentally acquired data. Received amplitudes are affected by the sensitivity of individual elements, variations of couplant on the surface of the sample, and roughness of surface finishes (particularly for manufactured defects). The anisotropy of CFRP can also result in variations in attenuation, which impact effect received amplitudes. Surface roughness and local changes in fibre density due to the materials inherent anisotropy produce small changes in time of flight. Traditional augmentation methods such as those used for images (e.g. crop, mix-up, flipping etc.) do not model these variations well and can produce unrealistic examples.

Therefore, in this study, we introduce two types of augmentation that were generated online for each minibatch during training. These augmentations aim to mimic the inter-element response variability observed within the UT probes used for data collection.

The first type of augmentation is related to the magnitude of response measured by the UT elements, which varies due to many factors not included in the simulation, such as manufacturing tolerances of the sample and the UT array probe, wear and tear of the probe and electrical wires/connections, or inter-layer multiple scattering of the sound waves. To mimic these whilst preserving the correct normalisation, each A-scan was scaled by a constant past the front wall. The scale factor was sampled from a uniform distribution to give a scale factor between 80-120%. An example of this is given in Figure 43 (a).

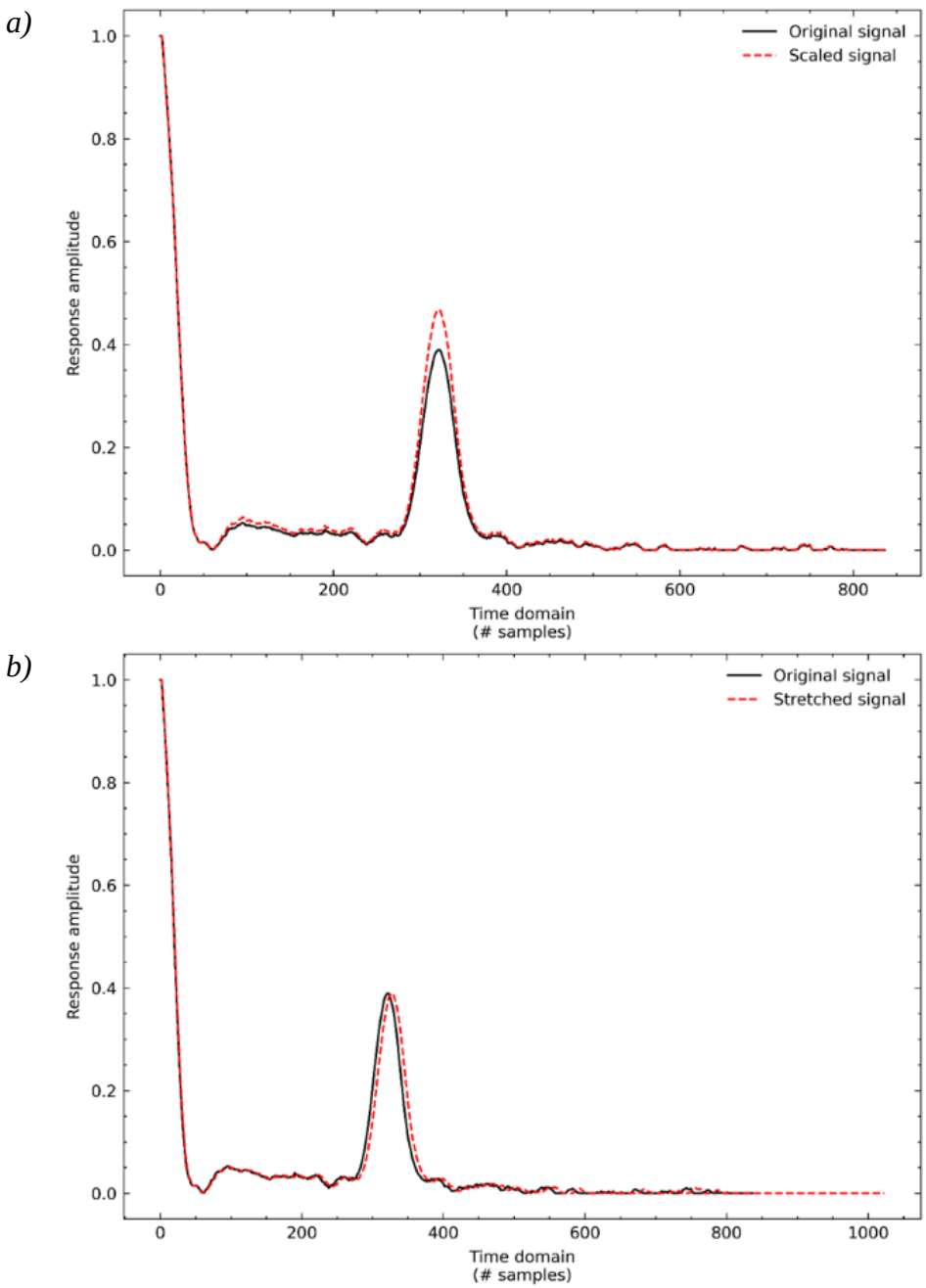


Figure 43: a) Example of how scaling augmentation is done on an individual A-scan. b) Example of how dilation augmentation and padding is completed for an individual A-scan.

The second type of augmentation mimics any changes in ultrasonic travel time seen by different elements. This can be caused by a variety of factors, such as variations in component sound speed due to the anisotropic nature of composites, departure from central frequency for certain elements, etc. To simulate this 1-D interpolation was used

to randomly stretch or compress the signal in the time domain. The dilation amount was randomly sampled from a uniform distribution for each A-scan up to $\pm$ 15 samples. An example of this is given in Figure 43 (b).

By introducing these augmentation methods, we aim to improve the generalisability of the models to the experimental domain. The online nature of these augmentations means that they can be easily incorporated into the training process without the need for additional data collection or pre-processing steps. To ensure consistent length of data in the time domain, each A-scan was padded with zeroes to a length of 1024 samples during training. Further investigation could be conducted to identify additional domain specific augmentation methods which could help to bridge the simulation to experimental domain gaps. This could include variations to frequency and bandwidth.

4.3 Network Architectures

In this work we investigated the performance of three different 3D CNN architectures for binary classification of 3D defect and defect free UT data with extreme aspect ratios.

The first 3D CNN, VoxNet, was designed for similar volumetric classification tasks and acts as a baseline architecture. For low aspect ratios CNNs (such as VoxNet) typically make use of square or cuboidal kernels which are appropriate for their equal (or near equal) aspect ratios. The use of CNNs on data with more extreme aspect ratios is less common and is particularly extreme for UT data between the time and the spatial domains, with an aspect ratio of 16.

To overcome this challenge, a task-specific architecture was hand-crafted by adapting VoxNet in a manner that follows the traditional approach to architecture design. This custom network is specifically designed to tackle the extreme aspect ratio problem and enhance overall classification performance.

As an alternative to traditional architecture design, neural architecture search was employed to develop a third architecture for comparison. For each model Adam optimiser [70] was used with a constant learning rate of 0.001, β_1 of 0.9 and β_2 of 0.999. A batch size of 8 was utilised in the training process. The chosen loss function for this model was binary cross-entropy, with a sigmoid activation function applied to the final layer to facilitate classification.

Due to the small amounts of experimental test data, there was a likelihood of noisy results during both training and testing phases. To mitigate this, each model was trained ten times with varying random initialisations, and their individual results were averaged across the performance metrics. This gives a better representation of the model's performance by averaging out any noisy results due to the small datasets.

During the training phase, a fixed validation set comprising 30% of the total test data was randomly selected from each class of experimental data. This set was used to monitor the model's performance and minimise the risk of overfitting. The models were trained with a patience of 10 epochs, where the training process monitored binary cross entropy loss on the validation data, for improvement. If there was no enhancement for a consecutive period of 10 epochs, the training process was halted. The model parameters with the lowest validation loss were used to evaluate the classification performance on the test set. This approach ensured that the final model's

defect detection performance was evaluated using the parameters that had the best ability to generalise to the target domain, as opposed to the model that had overfit to the synthetic domain.

4.3.1 Evaluation metrics

To quantitatively assess the binary classification performance of each network, average mean accuracy, precision, recall and F1 scores were calculated according to Equations 11-14.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (11)$$

$$Precision = TP / (TP + FP) \quad (12)$$

$$Recall = TP / (TP + FN) \quad (13)$$

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall) \quad (14)$$

Where *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative, with positives being the presence of a defect within an inspected volume. Each result was individually averaged using a simple mean across the 10 training cycles.

4.3.2 VoxNet: Baseline Architecture

Introduced by Maturana and Scherer, VoxNet [162] is a 3D CNN designed to tackle classification problems of 3D data that can be represented as voxels to form an occupancy grid. Originally tested on LiDAR, RGBD and CAD data it has since been used as the backbone for methods tested on ModelNet40 [163].

While the data from UT for this task differs from the datasets previously employed with VoxNet, the process of converting data into voxel-based format within the

VoxNet pipeline is well aligned to the 3D representation of UT data. As a result, VoxNet was employed to establish baseline model performance metrics for this task.

VoxNet is constructed using two 3D convolutional layers with cuboidal kernels, followed by a pooling layer and two fully connected layers (Figure 44). For further details on the model please refer to the original paper. VoxNets total number or parameters is 235M.

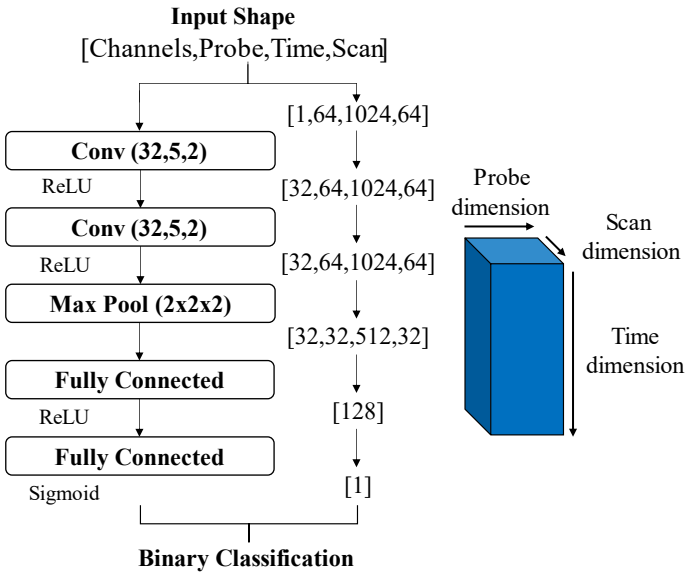


Figure 44: The VoxNet architecture. Where Conv (f,d,s) indicates the number of filters f, filter size d, and stride s, of the convolutional layer.

4.3.3 Hand Designed Architecture

The second architecture, referred to as CustomNet, demonstrates a conventional approach to architectural design. In this context, adaptations to VoxNet have been implemented to contemporise and enhance its performance specifically for the given task.

The UT dataset stands out to previous VoxNet datasets due to its higher dimensionality coupled with notable differences in spatial and temporal dimensions. To effectively

handle these unique attributes, adjustments were made to the models' architecture. Specifically, the number of convolutional layers was increased to enhance the extraction of meaningful features from the complex dataset. Additionally, cuboidal kernels with non-uniform dimensions were employed in the initial four blocks (refer to constant blocks in Figure 45) of the model. This approach aimed to address the uneven dimensionality inherent in the data, ultimately equalising the dimensions and contributing to a more robust feature representation throughout the network (Figure 45). After this a feature block with cube kernels of equal dimensionality could be used (refer to feature block in Figure 45).

In the process of updating VoxNet, we incorporated convolutional layers for pooling instead of the previously employed max-pooling layers. Additionally, ReLU was substituted with LeakyReLU, and batch normalisation was introduced. To mitigate overfitting, dropout and global average pooling were employed to reduce the number of features for classification, avoiding the use of large fully connected layers. These modifications are geared towards improving the model's performance by incorporating contemporary practices that have shown substantial performance benefits, as highlighted in previous studies [82].

The final architecture is given by the diagram in Figure 45. The total parameter size of the network was estimated to be 1.28 M parameters.

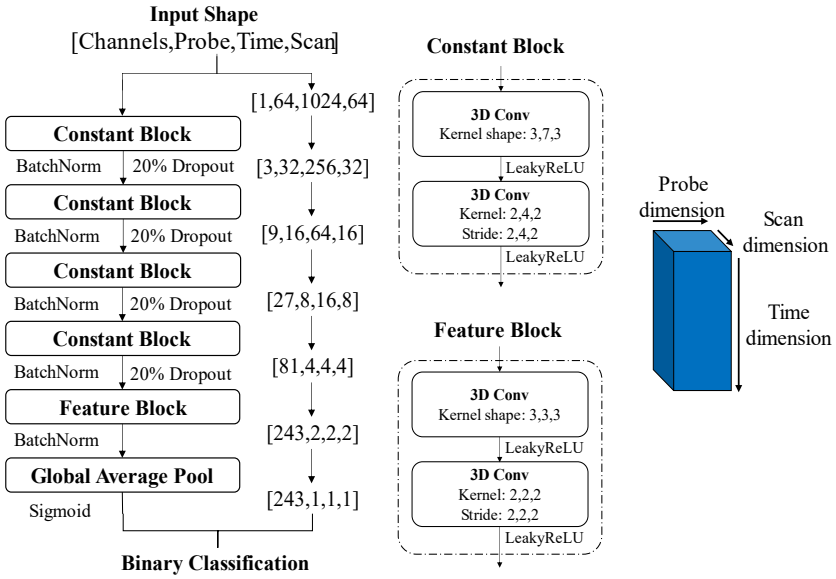


Figure 45: Network architecture for the CustomNet.

4.3.4 NAS Discovered: 3D ResNet based Neural Architecture Search

4.3.4.1 Neural Architecture Search

The final architecture was developed through NAS of a modified ResNet search space to account for 3D convolutions and operations. One of the challenges in applying NAS to a new domain task is the design of the search space. For this task, a new search-space framework which utilises a novel search space based on a ResNet-like structure is introduced. A fixed stem was used to down sample the data by a factor of 4 in the spatial dimensions and a factor of 8 in the time dimension whilst aiming to retain information through increasing the channels to 64. A further down sample block with average pooling followed by two to four residual blocks were all searched individually. An overview of the structure can be seen in Figure 46. The residual blocks and bottleneck features of the ResNet architecture are retained, whilst searching operations for each edge within the residual block. This provided a large diversity of architectures, which is key to attaining good performance in a novel application, whilst also ensuring

that many networks conformed to successful design principles. Each residual block contained two fixed point-wise convolutions used to down and up sample the number of channels. Figure 47 shows an example of a residual block denoting the searched and fixed operations.

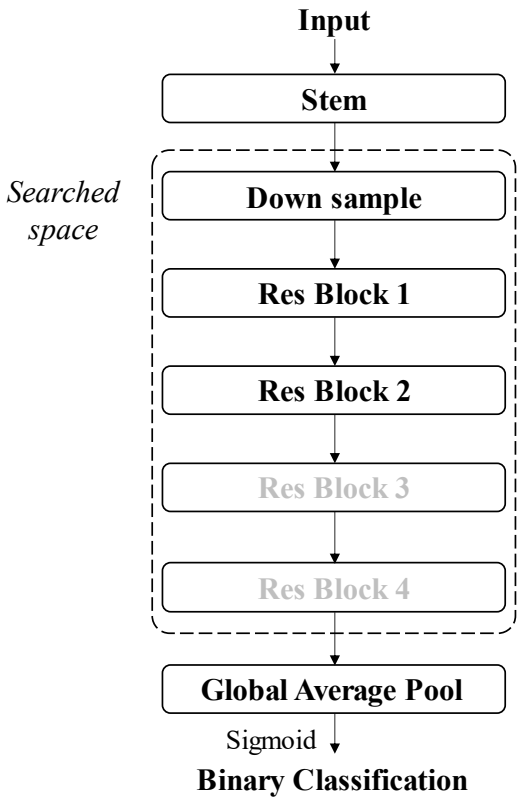


Figure 46: Representation of the ResNet style searched space.

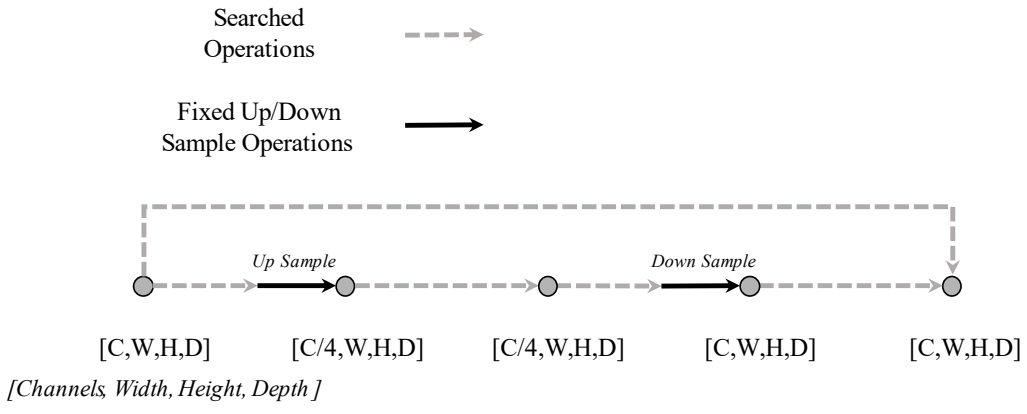


Figure 47: Diagram of the searched residual block.

These blocks were then stacked in groups, with the resolution down sampled between groups. Equation 15 gives the probability of a new group being created for each residual block, otherwise they were added to the current group. This makes groups unlikely to be extremely long or short.

$$P(newResBlockGroup) = \frac{1}{currentResBlockGroupSize} \quad (15)$$

The primitive operations of a search space are the list of operations which are assigned to the edges of a network architecture. The implemented approach incorporated a standard set of operations commonly found in the NAS literature. These operations comprised of convolutions, pooling, and skip connections, which are widely recognised and utilised within the field. These operations were all 3D due to the dimensionality of the data. In contrast to standard practice, which makes use of separable convolutions, the approach presented in this study deployed both depth-wise and point-wise convolutions as the fundamental convolutions within the search space. This significantly reduced the number of parameters in each operation of the architecture, greatly reducing the computational cost. Specifically, the depth-wise convolutions were applied with equidimensional cube kernels, of size 3, 5, or 7, coupled with dilation values that ranged from 1 to 4. Skip connections, point-wise convolutions, as well as average and max pooling operations were also searched for. For the pooling operations, equidimensional cube kernels of size 3, 5, or 7, with a dilation value of one were employed. The search encompassed the exploration of Gaussian Error Linear Unit (GELU) activation function and batch normalisation, as well as the absence of activation and normalisation operations. This allowed for architectures with fewer activation and normalisation function which has been shown

to be beneficial [82]. The searched down sample operation had a fixed kernel size with two in the spatial dimensions a four in the temporal dimension with a dilation of one. Throughout the relevant operations, a stride of one was employed.

A simple random search was applied to this search space for 80 iterations. Each model was evaluated using the validation dataset, with the lowest loss on validation across the training taken as the evaluation metric. For each searched architecture, a model was retrained with new initialisations three times and the mean evaluation metrics were used when selecting the discovered architecture, this ensured a more accurate estimate of model performance. Cross validation was unable to be used as the combination of NAS and domain transfer would have resulted in data leak between the NAS stage and the final model test evaluation stage. Figure 48 provides an overview of the NAS process and demonstrates how separation of the validation and test set were maintained in context of the complete model pipeline, given in Figure 40.

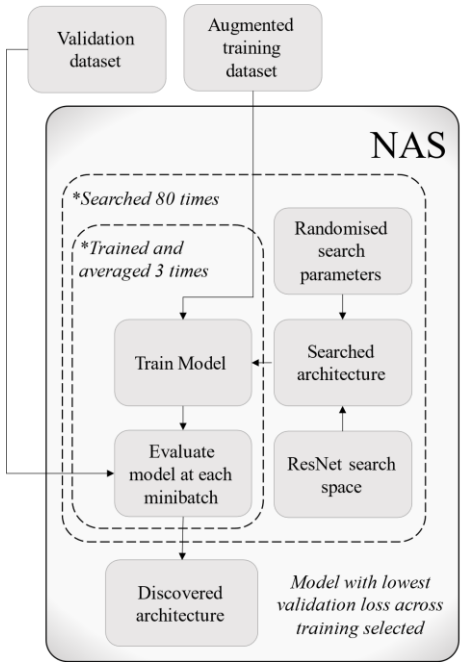


Figure 48: Overview of the process for NAS implementation.

The final discovered architecture had 1.03 M parameters and is given in Figure 49, with the details of the residual blocks given in Figure 50.

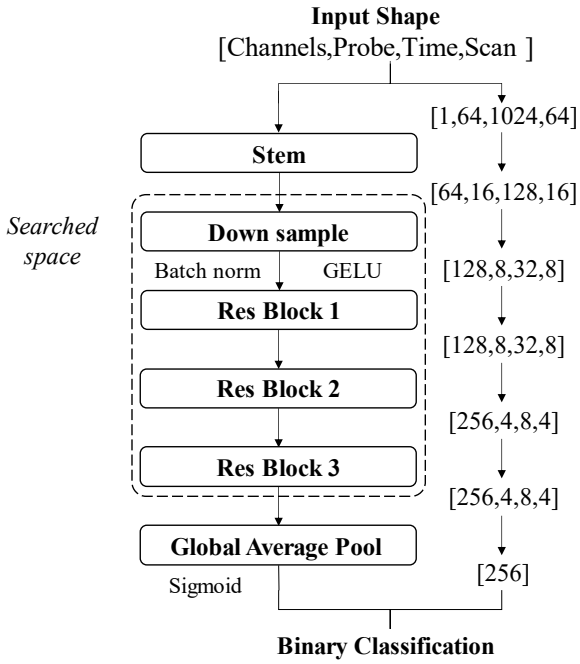
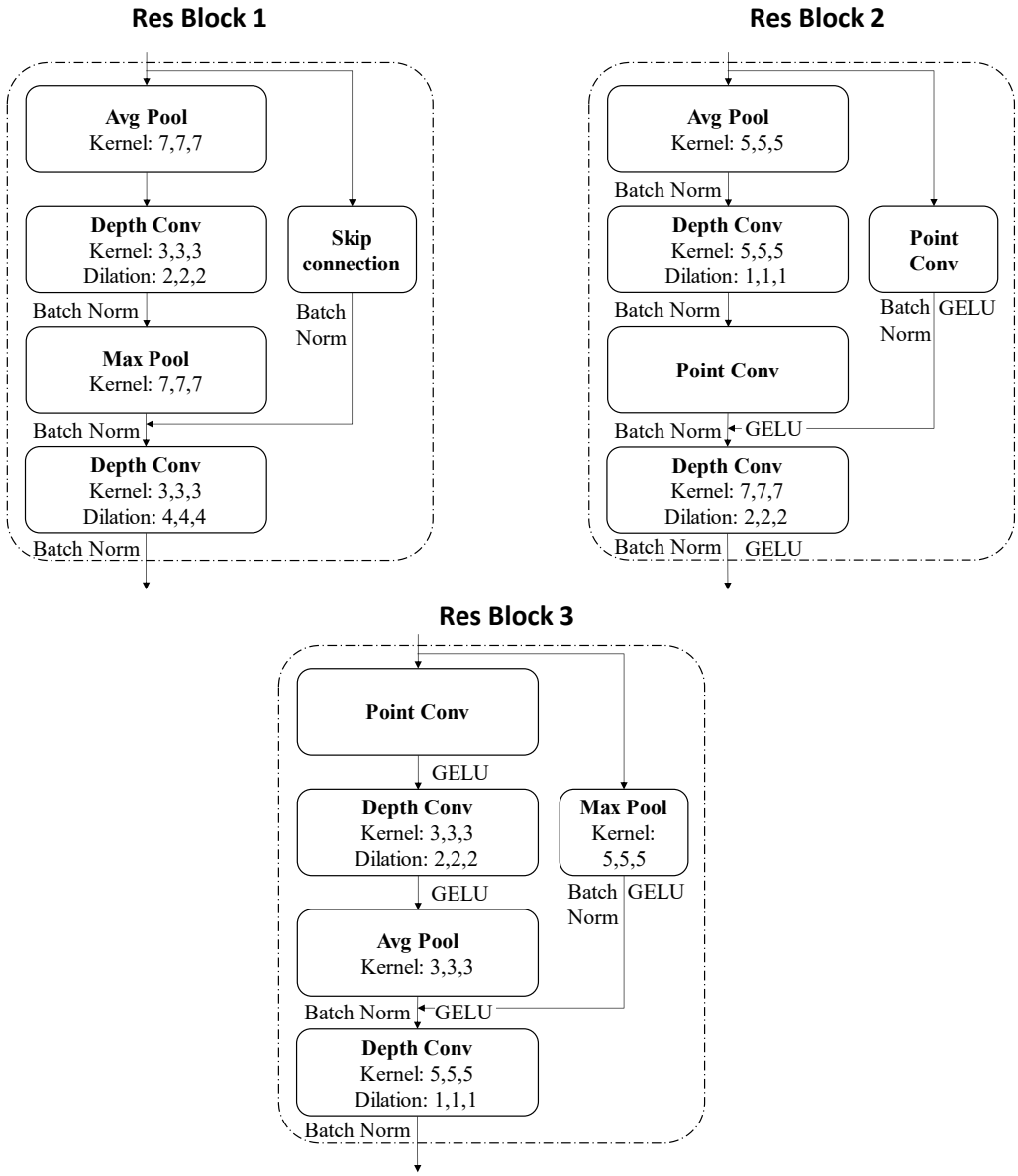


Figure 49: The overall structure of the discovered architecture.



1 *Figure 50: The details of each discovered residual block.*

2

Results

Table 15:Average confusion matrices for VoxNet, CustomNet and the NAS discovered architecture.

	True	Predicted	
		No Defect	**Defect**
VoxNet	**No Defect**	15.8	9.2
	Defect	2.1	22.9
CustomNet	**No Defect**	24.3	0.7
	Defect	3.2	21.8
NAS Discovered	**No Defect**	25.0	0.0
	Defect	0.0	25.0

Table 16: Comparison of classification results across the different architectures. The means and standard deviations are presented as mean ± std.

Model	**VoxNet**	**CustomNet**	**NAS**
Accuracy	0.774 ± 0.184	0.922 ± 0.095	**1.00** ± 0.00
F1	0.825 ± 0.114	0.904 ± 0.130	**1.00** ± 0.00
Precision	0.793 ± 0.219	0.975 ± 0.050	**1.00** ± 0.00
Recall	0.916 ± 0.073	0.872 ± 0.197	**1.00** ± 0.00

Table 15 provides average confusion matrixes for the test results of the VoxNet, CustomNet and NAS discovered models. Table 16 presents a summary of each method's performance, displaying the mean and standard deviation across various performance metrics.

The architecture discovered by NAS consistently produced ideal results when trained using data augmentation, with a mean classification accuracy of 1.00 and a standard deviation of 0.00 across the 10 separate training iterations, this demonstrated high confidence in the model’s conclusions and robust design for the target domain.

Table 17: Comparison of the effects of data augmentation on the NAS discovered architecture. The means and standard deviations are presented as mean ± std.

Augmentation	None	Scaling	Both
Accuracy	0.776 ± 0.178	0.806 ± 0.227	**1.00** ± 0.00
F1	0.830 ± 0.123	0.842 ± 0.168	**1.00** ± 0.00
Precision	0.745 ± 0.19	0.846 ± 0.227	**1.00** ± 0.00
Recall	0.972 ± 0.044	0.916 ± 0.182	**1.00** ± 0.00

Table 18: Comparison of model sizes and inference time for each architecture.

Model	VoxNet	CustomNet	NAS
Total Parameters (M)	235	1.28	**1.03**
Total Size (MB)	1779	557	**93**
Inference Time (seconds)	0.37	**0.03**	0.40

Table 17 demonstrates the impact of data augmentation on the best performing NAS model. Discarding data augmentation completely during training had a significant impact on the classification performance with a 22.4% drop in mean accuracy, along with a standard deviation increase in 17.8%, which demonstrated a significant reduction in statistical confidence. Whilst the addition of amplitude scaling augmentation improved the mean accuracy, it was only by 3%. This demonstrates the importance of using both augmentation methods in parallel for increased generalisability to the experimental domain.

Table 18 provides a summary of model sizes, and inference times for a single batch of test data. The NAS discovered architecture has a total size 16.7% and 5.2% smaller than CustomNet and VoxNet respectively. Whilst the CustomNet was 12 times faster at inference than the next closest, VoxNet.

4.4 Discussion

VoxNet demonstrated it was able to learn features from synthetic data and performed reasonably well on experimental data, with a mean F1 score of 0.825. However, its significant standard deviation in accuracies between training instances demonstrates that the architecture was not well optimised for the problem. The CustomNet improved on the accuracy of VoxNet substantially by 14.8%, whilst also reducing the standard deviation of results by 8.9% which indicated an increase in consistent generalisability to the experimental domain. This illustrates the benefits of tailoring architectural modifications to address the needs of specialised tasks. The experimental results demonstrated that the architecture discovered from NAS greatly outperformed the other two in terms of classification accuracy. Whilst all the models used in this work are not large and are considerably smaller than typical sizes for 2D ResNet's and other CNNs [156], the NAS model was able to achieve the highest performance with a significantly lower model size, at only 5.2% the memory requirement of VoxNet. The black box nature of DL makes it difficult to specify which features lead to this improvement in performance. This complexity is in fact a large motivator for NAS as the design space is too large for a human to efficiently find an optimal network architecture. It is anticipated that the addition of skip connections and the ability to vary operations at different depths added by the NAS has a significant positive impact on performance. This demonstrated the importance of utilising neural architecture search to optimise CNNs.

Due to the large fully connected layers, VoxNet results in a far greater number of total parameters than the other two networks. This results in a model which occupies far

more memory. Whilst CustomNet and the NAS discovered architecture have a comparable number of parameters the discovered network is far smaller. This is a result of many of its operations being far more efficient, such as the separation of point and depth wise convolutions. Whilst the discovered architecture is the smallest, its inference time takes the longest due to the greater architectural complexity of the model and its operations. This said, all models have acceptable inference time and can process 8 samples in under half a second. However, CustomNet is notably twelve times faster at inference than the second fastest network, VoxNet, which could be an advantage in some industrial settings.

When trained without data augmentation the NAS model performed significantly worse. Furthermore, the performance was only slightly improved by adding amplitude scaling augmentation alone. For best performance, both augmentation methods were needed in combination. This indicates that despite accurate synthetic data generation, data augmentation still has a significant role in producing generalisable models to the experimental domain.

Whilst ideal classification was achieved consistently for the discovered architecture when trained with data augmentation, this was tested on detection of manufactured defects only. Specifically, back drilled holes which are perpendicular to the propagating sound wave and act as ideal reflectors. This makes them comparably easier to detect than other defects. Whilst samples with naturally occurring defects are challenging to get access to, future work would benefit by expanding the simulation scope and testing the models on naturally occurring defects which will likely prove more challenging to detect. For more challenging detection and characterisation tasks

a more sophisticated search optimisation algorithm could be employed to discover architectures more efficiently.

The achieved classification results suggest that the synthetic data generation process is a viable approach for producing fully synthetic 3D UT volumetric datasets that closely map to the experimental domain and enable the development of effective classifiers. However, due to the substantial improvement in classification performance achieved through the implementation of data augmentation methods, it is important to acknowledge that disparities between the synthetic and experimental domains persist. This observation underscores the necessity for augmentation techniques to further enhance the generalisability of the model. Nonetheless, it is worth noting that the data augmentation methods employed in this study proved to be highly effective in aiding not only generalisability but also in facilitating the transfer of knowledge across domains.

The key benefits for analysing the complete 3-D volumetric data instead of processed images were the ability to learn from greater features, the reduction in pre-processing requirements, and the potential reduction in inference time by analysing the complete volume all at once. The impact on inference time is challenging to quantify, however if comparing the compute required to process 64 B-scan images (the equivalent spatial scan data), without parallelisation for equivalent 2D classifiers, there is the potential for up to 64 times saving in inference time for the same scan area. Despite these advantages there are still potential benefits to analysing UT data as images. One of these is the many opportunities for detection of a single defect in multiple B-scans. It is likely that defects will span multiple B-scan images, and as such by analysing each

B-scan there are multiple chances to detect an individual defect. This means an individual defect can still be detected even if individual defective images are incorrectly classified. However, the opportunities for characterisation and localisation of defects are far greater when retaining the volumetric spatial information and this work opens future prospects for 3D classification and segmentation which would be much more challenging if using C-scans or B-scans alone.

The research outcomes demonstrated the considerable potential of employing 3D-CNNs in conjunction with well-designed data augmentation techniques and optimised architecture search spaces to address challenging 3D classification tasks characterised by extreme aspect ratios, as observed in the context of UT. Insufficient utilisation of data augmentation severely hampered the model's ability to generalise to experimental datasets, leading to suboptimal classification performance. Likewise, choosing an unsuitable model architecture could result in the failure to capture crucial features necessary for accurate classification. Consequently, it is imperative to thoroughly consider both aspects during the design of a classification model for 3D UT data to ensure optimal performance.

4.5 Conclusion

Deep learning has demonstrated prior success in ultrasonic non-destructive evaluation when applied to either time series or image data. However, analysing only time series or image data can result in a significant loss of information in either the temporal or spatial domains. This work proposes the use of 3D convolutional neural networks to classify complete volumetric ultrasound data without compression, retaining all spatial and temporal information. This approach not only reduced the need for accurate gating when constructing C-scan images but also decreased the amount of signal processing

required. To train the models, synthetic data was generated from semi-analytical simulations, while experimentally collected ultrasonic responses from manufactured defects were used for testing. Two forms of data augmentation were implemented based on physical variations seen in experimental ultrasonic responses to improve the model's classification performance in the experimental domain. Furthermore, the performance of three different architectures; one existing in the literature, one hand-designed based on current practices, and one designed by NAS from a ResNet search space modified for 3D, were compared.

The first architecture, VoxNet, performed reasonably well on experimental data, achieving a mean F1 score of 0.825. However, its notable standard deviation in accuracies during training suggests suboptimal architecture optimisation for this task. CustomNet's greatly improved on VoxNet with an accuracy increase of 14.8%, whilst reducing the standard deviations in accuracy by 8.9%, hence demonstrating an architecture better optimised for this task.

The third architecture, designed by NAS, when trained with data augmentation, gave the best results, providing 100% classification accuracy. The impact of online domain specific augmentation was notable, leading to a 22.4% decrease in mean accuracy for the NAS model when augmentation was omitted.

Overall, this work demonstrated that it is possible to train successful DL models to classify full volumetric ultrasonic data for NDE. The issue of a lack of data in most NDE situations was addressed by successfully implementing synthetic data generation in 3D. The work highlighted the importance of appropriate architecture selection and

effective data augmentation when translating between synthetic and experimental domains, with both factors essential in achieving high classification accuracy.

The focus of this work was on the use of volumetric datasets, and whilst 100% classification accuracy was achieved through effective NAS, it is recognised that FBHs are generally simple defects to detect by human operators.

Future work aims to increase the complexity of the task by detecting a wider range of more challenging defects and expanding the simulation scope to better cover naturally occurring defects, where performance can be measured against human operators in a more realistic industrial scenario. Further work also aims to extend the problem to defect classification and sizing.

5 Supervised Volumetric Defect Segmentation

5.1 Introduction

Automated NDE data interpretation and reporting can be broken down into distinct sections which in many cases will flow sequentially, as presented in Figure 9. In most settings, defect detection is the primary task to be completed. Once defects are detected it is important to evaluate their size against acceptance criteria. In some applications the acceptance criteria will vary depending on the defect type, orientation, and location, so the identification of different defect types and the location in the geometry is also important. In an industrial setting it is crucial to report on the NDE findings for traceability and to allow for downstream testing, design decisions, and re-work. It is therefore not acceptable to just detect a defect. Defects must be evaluated to extract as much information about the defect as possible, such as sizing, type, location etc [165]. This is covered by the characterisation and quantification section of the pipeline (Figure 9). Whilst there is potential for a single end-to-end model/system to complete the whole interpretation process; by breaking the tasks up it will allow for greater model testing and improved understanding and comparison to human operators. Breaking down the data interpretation process into separate tasks not only enables comprehensive model testing but also provides valuable insights into how errors propagate through the system. This decomposition allows for a more nuanced evaluation, shedding light on the strengths and limitations of each component of the analysis. Additionally, the isolation of specific tasks enhances the ability to discern how errors manifest at different stages, making it possible to facilitate a more effective comparison and agreement with human operators. Their expertise can be leveraged to refine and optimise each element. This stepwise methodology ensures a thorough

assessment of the entire system and establishes a foundation for collaborative decision-making between automated systems and human operators. Moreover, this modular approach enhances the system's adaptability to different requirements, allowing for customisation and optimisation based on specific application needs.

Much of NDE automation research focuses on defect detection, often overlooking the significant burden placed on NDE operators for report generation. By analysing complete ultrasonic volumes of a sample, this work focuses on defect sizing and 3D localisation, which can be used directly when evaluating defects against accept/reject criteria and provide accurate defect positional information, which is useful in supporting re-work etc. This is accomplished through segmentation of ultrasonic volumes, which allows for automated generation of computer-aided design files to further reduce the burden on NDE reporting and can be used for building digital twins of components for testing [166], [167].

Accurate defect sizing is a key metric in determining if a component is safe against standardised acceptance criteria. The 6 dB drop method is a widely accepted method for defect sizing and is commonly used in industrial standards [16]. The technique relies on the utilisation of a single transducer and the peak amplitude from the defect response to determine the boundaries of a defect response by detecting the point at which the transducer is directly over the edge of the defect as determined by a 50% energy dissipation from the reflector, manifesting as a 6 dB reduction in amplitude as indicated in Figure 51. The method benefits from being based on physical properties and is fully explainable. The 6 dB drop is often extended and applied to amplitude C-Scans and phased arrays for thresholding defect areas [92], [168], [169]. Whilst the 6

dB drop method is widely established it does have limitations. Primarily, the defect must be larger than the acoustic beam to get an accurate value for the peak response amplitude [168]. In addition, real defect responses generally do not follow the ideal defect response curve, often leading to under-sizing defects [170]. To combat this, alternative amplitude drop thresholds are used in different industrial settings [16], [171]. These are often component specific and require experimental determination. Li et al. proposed an alternative method which utilised a generalised regression neural network and took additional features into account to provide dynamic thresholding of a C-scan image for more accurate defect sizing than the 6 dB drop method [172].

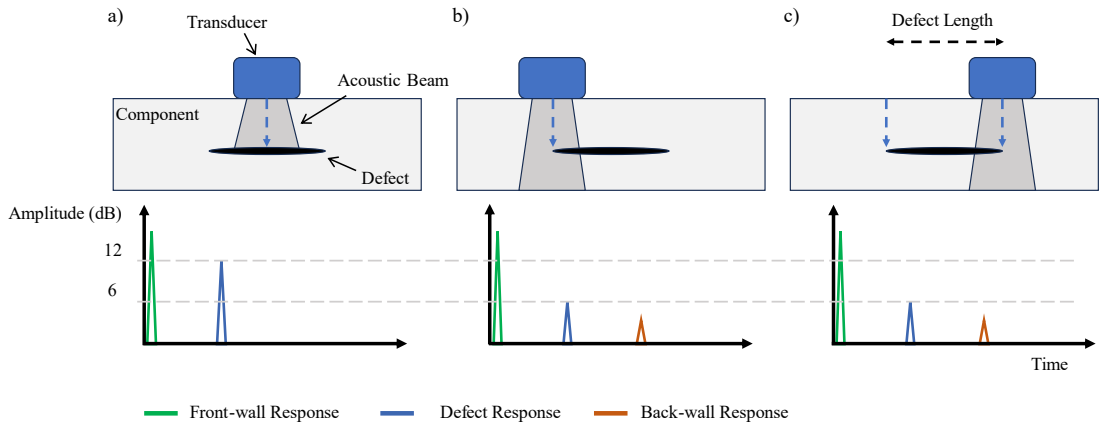


Figure 51: Demonstration of the 6 dB drop method for defect sizing. a) Finding maximum defect response. b) Using the 6 dB loss in maximum amplitude to locate one edge of the defect. c) The corresponding defect edge detected using the 6 dB drop to determine the defect length.

Whilst the 6 dB drop can be used for in-plane defect localisation, depth-wise localisation requires information from the time trace signal. Cheng et al. showed a promising method for depth localisation of defects in CFRP panels using different DL approaches [173] with A-scan signals. They reported a minimum depth relative error of 9% for the hybrid CNN-LSTM, reducing the error by 96% compared to relying on the peak-to-peak time-of-flight measurement alone. DL models present an avenue for

advancing the automation of NDE data interpretation, and are becoming more prevalent in the literature, especially in the context of defect detection when dealing with images or A-scans [14]. DL is particularly well suited to addressing challenging automation tasks, where a traditional method may not be available, such as defect characterisation and quantification. There are several examples of DL models demonstrating the ability to exceed human performance in certain situations [100], [174].

The previous chapter demonstrated the effectiveness of using DL to detect defects from volumetric ultrasonic data. This chapter presents an alternative method for defect sizing using 3D U-Net for volumetric segmentation of ultrasonic data, evaluated against the established 6 dB drop method. U-Net is a DL model, introduced in 2015, which proposed an architecture for medical 2D image segmentation [175]. The authors highlighted the problem that for localisation and classification, pixel level annotation is required which is much more intensive. This makes large, labelled, datasets unreachable. They highlight that the sliding window approach which is often used suffers from two key drawbacks: 1. It is inefficient to run the network for each window, and 2. There is a trade-off between localisation accuracy and use of context. Despite advances in computer vision the U-Net architecture is still widely popular and shows impressive results in many different segmentation tasks [102]. Çiçek et al. [176] extended the original U-Net paper for 3D segmentation of highly variable kidney volumes, giving 3D U-Net, which showed impressive results. By incorporating algorithms with the capability to interpret volumetric data, it is ensured that all spatial and depth wise information is preserved. This approach provides the model with more pertinent features to learn from and eliminates the necessity for image pre-processing

and gating. Furthermore, it enables comprehensive 3D defect localisation, a key focus of this study.

Section 5.2 provides information on the generation of synthetic training data, and the creation of ground truth segmentation masks. It also covers augmentation applied during training. In Section 5.3, the network architecture, training specifics, and the reference 6 dB drop method for sizing are detailed. The results and discussion are presented in Section 5.4, which is divided into localisation and sizing. To the best of the authors' knowledge, this marks the first utilisation of a 3D U-Net for sizing and localisation of defects in volumetric ultrasonic testing data, offering several advantages:

- Reduced pre-processing times as no thresholding, gating, or generation of images is needed.
- Results of the developed model trained exclusively on synthetic datasets, outperform industry standard 6 dB drop method for sizing by 35% on experimental test data.
- Complete localisation of defects within 3D space. Which enables easy extraction of data for downstream processes i.e. testing with Finite Element Analysis.

5.2 Data

The previous chapter utilised realistic synthetic volumetric data for training. The same synthetic training data was used to train the 3D U-Net model in a fully supervised manner without the need for any experimental training data. The trained model’s performance was evaluated against a fully experimental test dataset.

5.2.1 Mask Generation

An important advantage of employing simulated data as the foundation for training datasets lies in the capacity to have full control over the simulation input parameters. This control can be harnessed for the automated generation of ground truth masks during the training process, a task that would pose more significant challenges when training models on experimental data. In the context of this study, defect diameter and depths were utilised to create segmentation masks for defects with a nominal thickness. Figure 52 provides an illustration of a ground truth defect mask alongside its corresponding simulation.

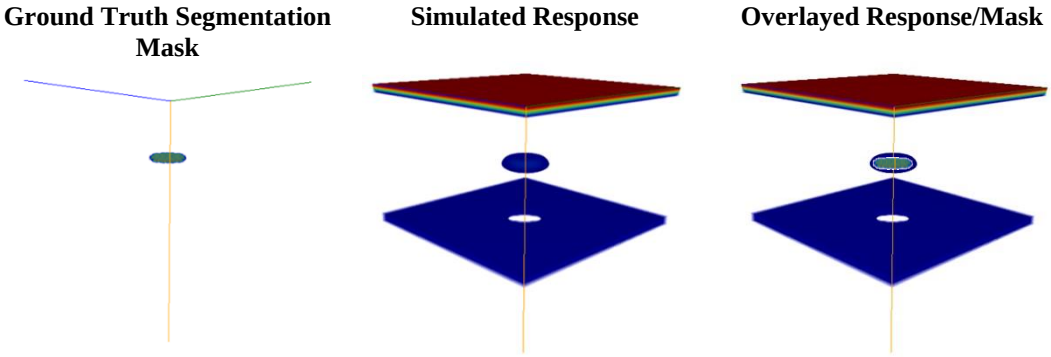


Figure 52: (a) The ground truth segmentation mask and (b) the corresponding simulated defect response. The overlay of both the mask and response is shown in (c). Colour mapping and axes are given in Figure 21.

Table 19 provides a summarised description of the datasets created from both the experimental and synthetic data sources.

Table 19: Summary of the datasets produced.

Data source	Dataset	Number of datapoints
Simulated defect responses and segmentation masks (300 Flat-Bottom Holes)	Synthetic defective || T*rain (80%)*	240
	Synthetic defective || V*alidation (20%)*	60
Experimental defect sample (25+15 Flat-Bottom Holes)	Sample 1 ~ Diameters: 3, 4, 6, 7, 9 mm || *Test*	25
	Sample 2 ~ Diameters: 3, 6, 9 mm || *Test*	15

5.2.2 Augmentation

The generalisability of ML models is a critical aspect of their performance and accounts for differences in the source and target domain. Augmenting the training data improves model generalisability by adding noise at the training stage, reducing the likelihood of overfitting. This often improves performance in the target domain particularly when the target (experimental) domain is different from the source (synthetic) domain.

In this study, two domain-specific augmentation techniques that have shown their effectiveness in the previous chapter for augmenting volumetric ultrasonic responses were employed during training. Standard computer vision augmentation methods (e.g. mix-up, cut mix etc.) do not translate directly to UT data as they would impact the underlying signal response. The first type of augmentation is concerned with response magnitude. Magnitude can vary due to various factors unaccounted for in the simulation, such as manufacturing variances in the sample and the UT array probe, wear on the probe and its electrical connections, and the complexities of sound wave scattering between layers. To replicate these variations while maintaining the appropriate data normalisation, the amplitude of each A-scan was adjusted by a constant factor beyond the front wall. This factor was randomly selected from a uniform distribution, resulting in a scaling factor ranging from 80% to 120%.

The second augmentation method aims to replicate phase aberration - the variations in ultrasonic travel time between elements [177]. These variations can result from a range of factors, including fluctuations in the sound speed of composite materials due to their anisotropic properties, and deviations from the central frequency for specific elements.

To simulate phase aberration, a 1-D interpolation technique was employed to randomly stretch or compress the signal in the time domain. The extent of dilation was randomly determined from a uniform distribution for each A-scan, allowing for dilation of up to ±300 ns.

The objective in implementing these augmentation methods is to enhance the models' ability to generalise effectively within the experimental domain. The convenience of real-time augmentations allows for their integration into the training process, eliminating the need for additional data collection or preprocessing steps. To maintain consistent data length in the time domain, each A-scan was extended by zero padding, resulting in a length of 1024 samples during training. Subsequently, to mitigate computational demands, each volume was down sampled in the time domain by a factor of 4.

5.3 Segmentation Methods

5.3.1 Model: Architecture and Training

In this chapter, volumetric segmentation was carried out through the training and deployment of a customised 3D U-Net architecture. The design of the architecture drew inspiration from [176], but extended to five convolutional blocks with a sigmoid layer applied to the output. Models of varying number of blocks were tested, and the inclusion of an additional convolutional block resulted in a 40.8% reduction in validation loss. Further optimisation of hyperparameters and architecture may result in a performance increase, however this was outside the scope of this work. A graphical representation of the overarching architectural design is presented in Figure

53.

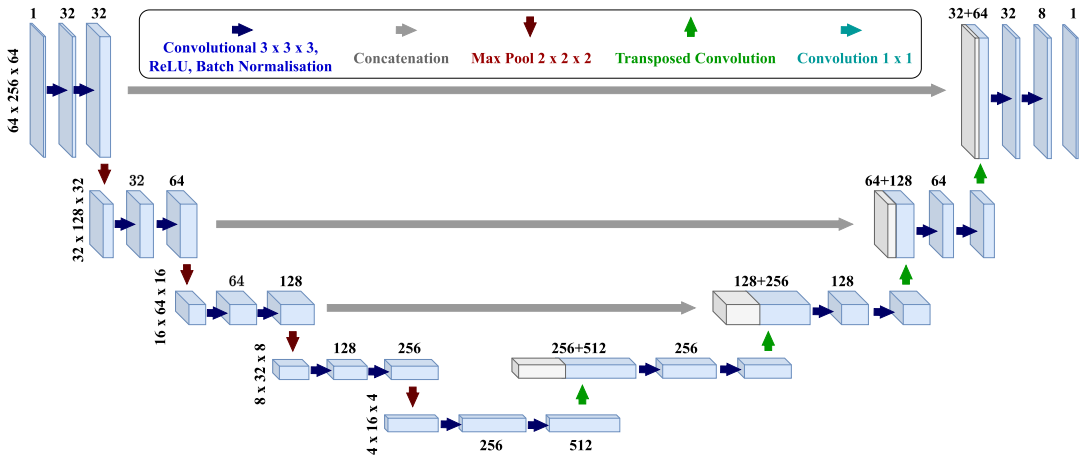


Figure 53: Architecture diagram for the 3D U-Net. Blue boxes depict the feature maps. The number of channels and dimensions of the data (probe × scan × time) are denoted above and to the side of each feature map respectively.

The model was trained on synthetic datasets with the corresponding defect masks generated from the previously defined parameter space (Figure 52). The model was trained with a batch size of 8 with a patience of 3 epochs. adopting a learning rate of 0.003. During training, binary cross-entropy loss was employed in conjunction with an Adam optimiser [70], which was parameterised with the default parameters given in the original paper (β1 set to 0.9 and β2 to 0.999) and a learning rate of 0.003. Training and testing of the models were conducted using a desktop Windows 11 PC with Nvidia RTX 3090 Ti GPU, 128 GB RAM, and two Intel® Xeon® Gold 6428 2.50 GHz CPUs. The PyTorch 1.13.1 library and Python 3.10.8 were used for coding. Overall training took less than 20 hours. For the subsequent testing phase, the model exhibiting the lowest validation loss at the 47th epoch.

5.3.2 Reference Sizing Metric: 6 dB Drop

The 6 dB drop criterion represents the prevailing industrial methodology for defect sizing, as documented in the literature [92], [169]. The underlying principle of this method relies on the utilisation of a single transducer to pinpoint the edge of a defect response by detecting the moment when exactly 50% of the energy is reflected by the defect, corresponding to a 6 dB decrease in amplitude [16]. The technique is repeated on the opposing boundary of the defect, and the resultant displacement results in the measured defect's length Figure 51. This fundamental principle can be further extended to encompass the sizing of defects from amplitude C-scan images produced from employing phased array transducers [92], [168], [169].

In this research, sizing in the fibre plane using the segmentations predicted by the U-Net model are compared with the established 6 dB drop. Given the prior knowledge of the manufactured defects present in the reference sample, recognised as circular FBH, and to mitigate variations in diameters through the component thickness, the defect diameters were computed based on the maximum segmented area through the depth of the sample, based on pixel summation, using the formula outlined in equation 16.

$$Diameter = 2\sqrt{\frac{\sum_{pixels}\left(\underset{through\ depth}{\text{Max}}(Segmented\ Volume)\right)}{\pi}} \qquad 16$$

5.4 Results and Discussion

Figure 54 presents examples of experimental defect responses and their corresponding segmentation masks as generated from the 3D U-Net. Along with demonstrating results, these visualisations could be used by human operators to sense-check the

models' predictions and reject inaccurate model predictions easily, providing the possibility for the method to contribute to a human-in-the-loop semi-autonomous NDE system.

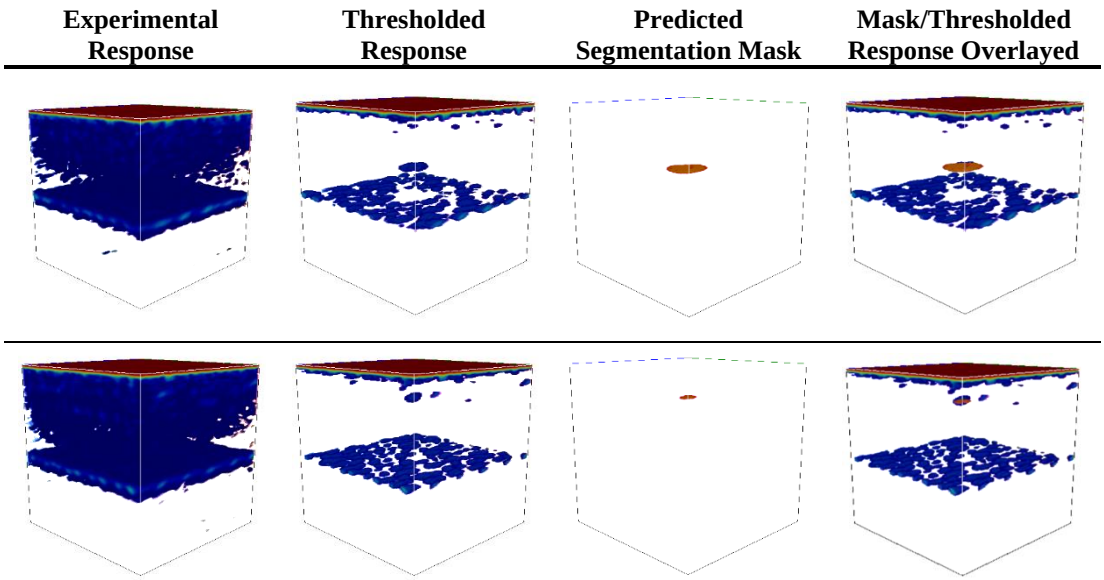


Figure 54: Example 9 and 3 mm defects respectively; their experimental ultrasonic volumetric responses, thresholded responses (amplitudes >10% of maximum response for defect visualisation), and their corresponding predicted segmentations. Colour mapping and axes are given in Figure 21.

5.4.1 Localisation

The localisation of defects in 3D can be deconstructed into two primary components: in-plane localisation and through-thickness depth-wise localisation. It is important to note that the widely adopted 6 dB drop criterion only addresses in-plane localisation and it does not provide information regarding depth-wise localisation.

The 6 dB drop method can produce inaccuracies in defect sizing as discussed in section 5.4.2. However, the circular shape of the test defects ensures that any errors in sizing, which might cause changes in diameter, will have minimal impact on the position of the defect's centroid. Obtaining an accurate ground truth for in-plane localisation less than 1.0 mm is infeasible due to the cumulative positional errors introduced throughout

the experimental setup. Therefore, this research uses the 6 dB drop criterion as a reference standard, to validate the agreement between the 6 dB and U-Net's in-plane localisation. By comparison, experimental through-thickness depth measurements for defects are considerably easier to acquire, which allows for a direct assessment.

5.4.1.1 Depth

The segmentation of volumetric ultrasonic data offers a distinct advantage compared to the 6 dB drop method due to its capacity for depth-wise localisation. This eliminates the need to employ multiple data types for characterisation, such as amplitude and time-of-flight C-scans. The depth-wise position determined from the segmented volume is represented by the mean segmented depth. In Figure 55 the predicted defect depth is compared to the true measured depth of the reference defects. The segmented volumes demonstrate a excellent level of accuracy in depth-wise localisation, as evidenced by a Mean Absolute Error (MAE) of 0.08 mm. This precision can be attributed to the substantially higher sampling rate in the temporal domain in contrast to the spatial domains, resulting in significantly superior temporal resolution when compared to spatial resolution.

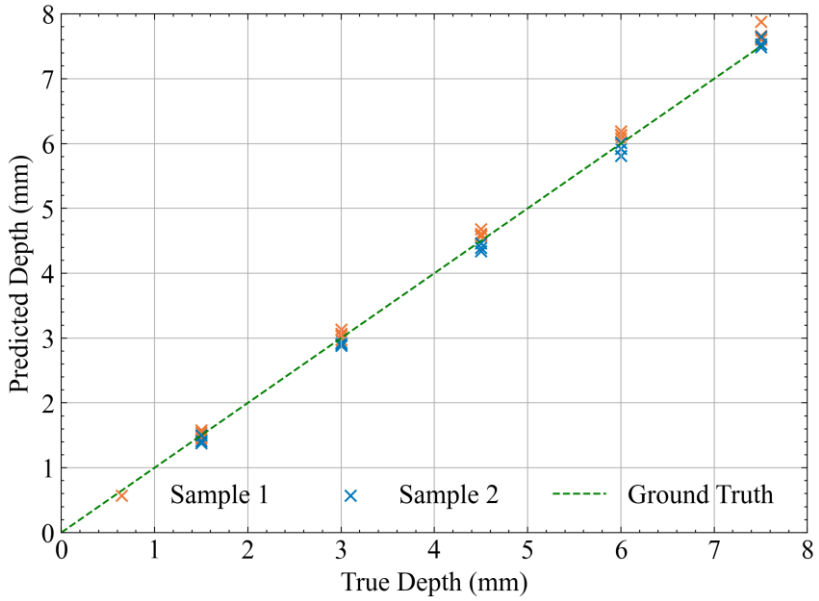
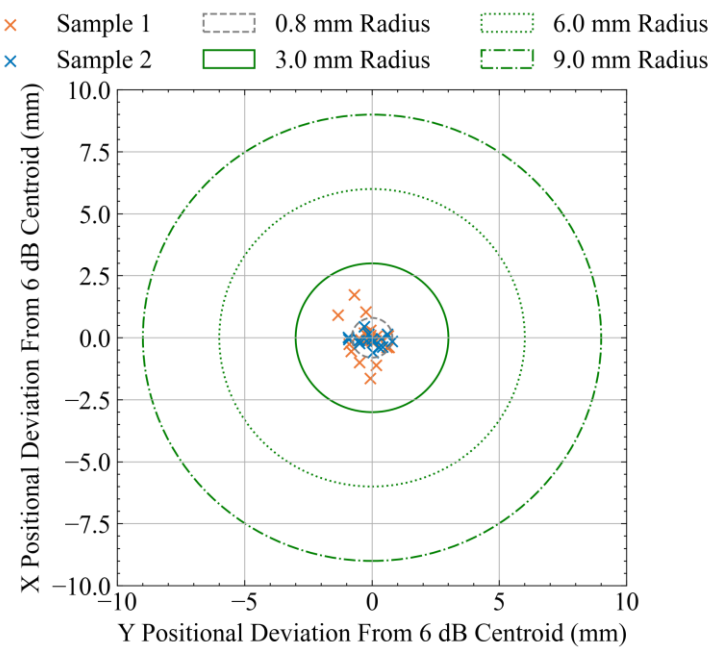


Figure 55: Depth localisation results.

5.4.1.2 In plane

The performance of the models' in-plane localisation is quantified by measuring their deviation from the centroid area compared to the 6 dB drop. Figure 56 visually presents the centroid deviation with reference to defect sizes and the pitch of array elements. As depicted in the figure, 75% of the variations (30 out of 40 defects) are below the 0.8 mm array pitch.

a)



b)

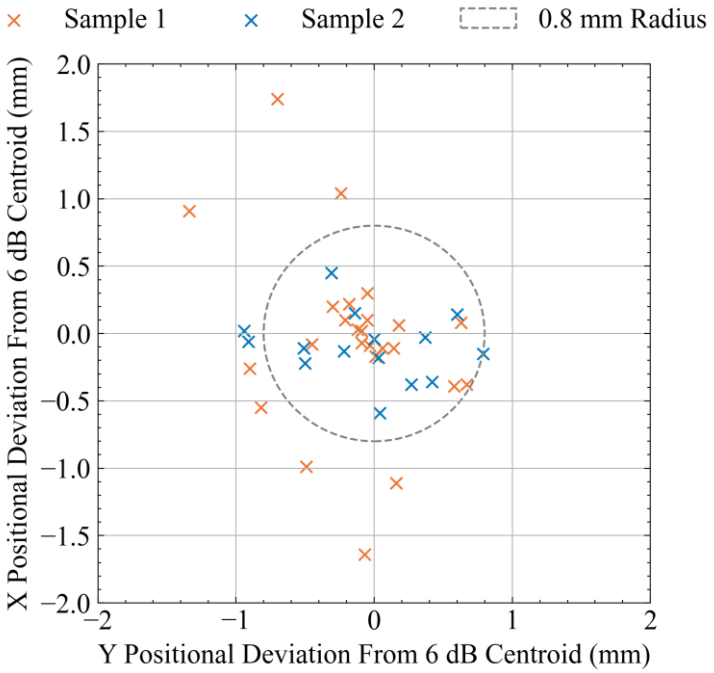


Figure 56: In plane localisation results compared to 6 dB drop with reference to defect diameters (a) and the expanded (b), which shows the reference to the array pitch more clearly.

Table 20 provides a comprehensive overview of the in-plane localisation outcomes, including the absolute distance between the 6 dB criterion and the centroid determined by the model. The Mean Absolute Error (MAE) of 0.57 mm demonstrates a substantial concordance with the established industrial benchmark represented by the 6 dB drop

criterion for in-plane localisation. Notably, the MAE, being less than the 0.8 mm array pitch, which establishes the spatial resolution, underscores the robust agreement between the model-based in-plane localisation and the standard reference.

Table 20: Complete model centroid deviation results from the 6 dB drop.

Sample 1				Sample 2			
Defect Diameter (mm)	**Deviation from X (mm)**	**Deviation from Y (mm)**	**Absolute displacement (mm)**	**Defect Diameter (mm)**	**Deviation from X (mm)**	**Deviation from Y (mm)**	**Absolute displacement (mm)**
	-0.05	0.10	0.11		0.42	-0.36	0.55
	0.58	-0.39	0.70		-0.50	-0.22	0.54
9	-0.49	-0.99	1.10	**9**	-0.94	0.02	0.94
	0.18	0.06	0.19		-0.31	0.45	0.55
	0.01	-0.17	0.18		-0.14	0.15	0.21
	0.63	0.08	0.63		0.37	-0.03	0.37
	-0.12	0.03	0.13		-0.51	-0.11	0.52
7	-0.07	-1.64	1.64	**6**	-0.91	-0.06	0.91
	-0.18	0.22	0.29		0.00	-0.04	0.04
	-0.09	0.02	0.09		-0.22	-0.13	0.25
	-0.03	-0.09	0.09		0.27	-0.38	0.47
	0.16	-1.11	1.12		0.04	-0.59	0.59
6	0.67	-0.38	0.77	**3**	0.79	-0.15	0.80
	-0.24	1.04	1.06		0.03	-0.18	0.18
	-0.09	-0.07	0.12		0.60	0.14	0.61
	-1.34	0.91	1.62				
	-0.45	-0.08	0.46				
4	-0.05	0.30	0.30				
	-0.70	1.74	1.88				
	-0.90	-0.26	0.94				
	-0.82	-0.55	0.98				
	0.14	-0.11	0.18				
3	-0.21	0.10	0.23				
	-0.30	0.20	0.36				
	0.06	-0.11	0.13				
Mean Average Error (MAE)			0.61				0.50
Total MAE							**0.57**

5.4.2 Sizing

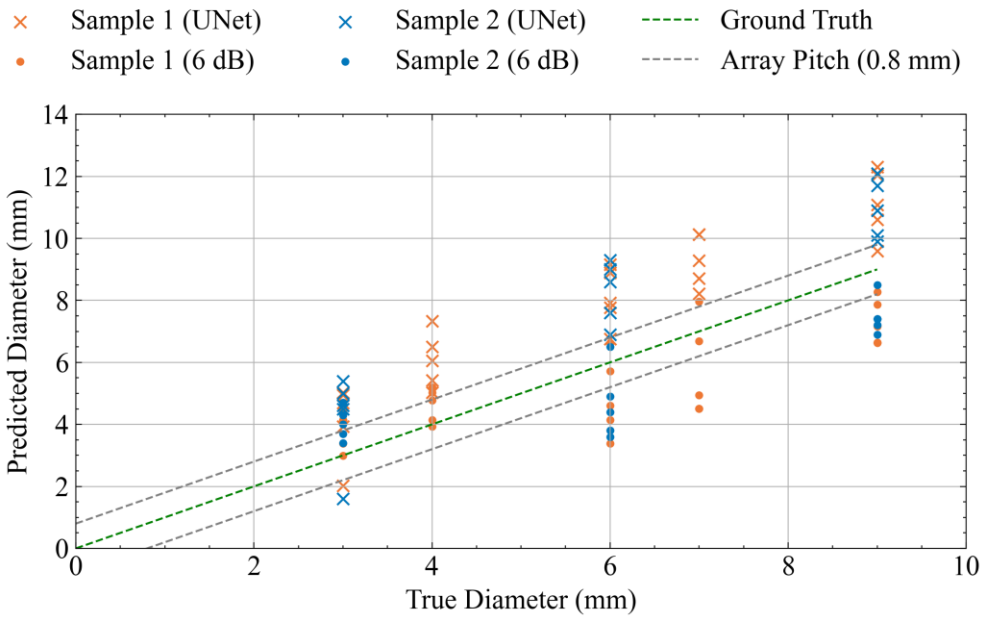


Figure 57: Sizing results for the 6 dB drop method and U-Net predictions.

Figure 57 presents a summary of the defect diameter predictions from the original segmented from the 6 dB drop and U-Net areas for each defect compared to the known ground truth area.

5.4.2.1 6 dB approach

With a MAE of 1.35 mm, our findings demonstrate a reasonable degree of accuracy, which, when coupled with suitable safety factors, is likely to be deemed adequate in industrial settings. Nevertheless, it is crucial to acknowledge that real-world responses often deviate significantly from the ideal defect response, leading the 6 dB drop approach to systematically underestimate defect sizes [170].

This limitation has prompted the utilisation of alternative amplitude drop methods for sizing in industry [16], [171], wherein the threshold values are frequently determined through experimental calibration. The experimental data collected for this research corroborates this tendency for under sizing defect responses (Figure 57), particularly

for defects exceeding 4.0 mm in diameter. Which exhibit a mean undersize of 1.37 mm. Conversely, our results reveal a tendency to oversize defects of 3.0 and 4.0 mm in diameter.

It is essential to recognise that any sizing method relying on maximum amplitude necessitates the defect to be substantially larger than the acoustic beam to accurately ascertain the maximum acoustic response. The experimental setup employed in this work, which utilised a 4-element sub-aperture with an element pitch of 0.8 mm, resulted in an effective transducer width of 3.2 mm. Considering this in combination with the spatial resolution limitations imposed by the fixed 0.8 mm pitch for each beam step, along with accounting for any beam spread, it was determined that the experimental setup was inadequate for the precise sizing of defects measuring 4 mm or less when employing an amplitude drop method. This leads to the average oversizing of 3.0 and 4.0 mm defects of 0.82 mm, counter to the expectation of the 6 dB amplitude drop under sizing defects. The under-sizing of defects larger than the diameter of the acoustic beam can likely be attributed to the curved edges of the defects. Since these defects do not maintain orthogonality to the sizing axis, the result is a diminished reflector when accounting for the three-dimensional nature of the response. This leads to a 6 dB decrease in acoustic energy closer to the centre of the defect, rather than at the true defect edge. The inconsistency of this method and the need for varying amplitude drop thresholds adds to the complexity of consistent defect sizing in industry and could be a concern for safety critical parts.

5.4.2.2 U-Net

The initial segmentation of U-Net masks yielded a MAE of 2.09 mm, which represents a 55% increase in error compared to the 6 dB drop method. As depicted in Figure 57, there was a consistent tendency to overestimate defect sizes across the range of diameters. This observation underscores the U-Net approach's reduced reliance on absolute peak amplitudes and its ability to deliver more consistent performance across a range of defect sizes, even when the defect size is not greater than the width of the acoustic beam. Moreover, it is worth noting that in numerous industrial applications for safety-critical components, it is preferable to overestimate rather than underestimate defect sizes.

It is imperative to delve into the reasons for the model's consistent trend of oversizing defects. Given that the model exhibited convergence during training on synthetic datasets, the overestimation observed during testing hints at a domain disparity between training on synthetic data and testing on experimental data. To elucidate this distinction, a comparative analysis using the 6 dB drop method was conducted between the responses derived from the experimental data and those generated by synthetic data for corresponding defect sizes and depths, as illustrated in Figure 58 (a). As previously noted, defects of 3 and 4 mm in diameter were too small to be accurately sized using this experimental setup and the 6 dB drop method, and thus, were excluded from this analysis (Figure 58 (b)).

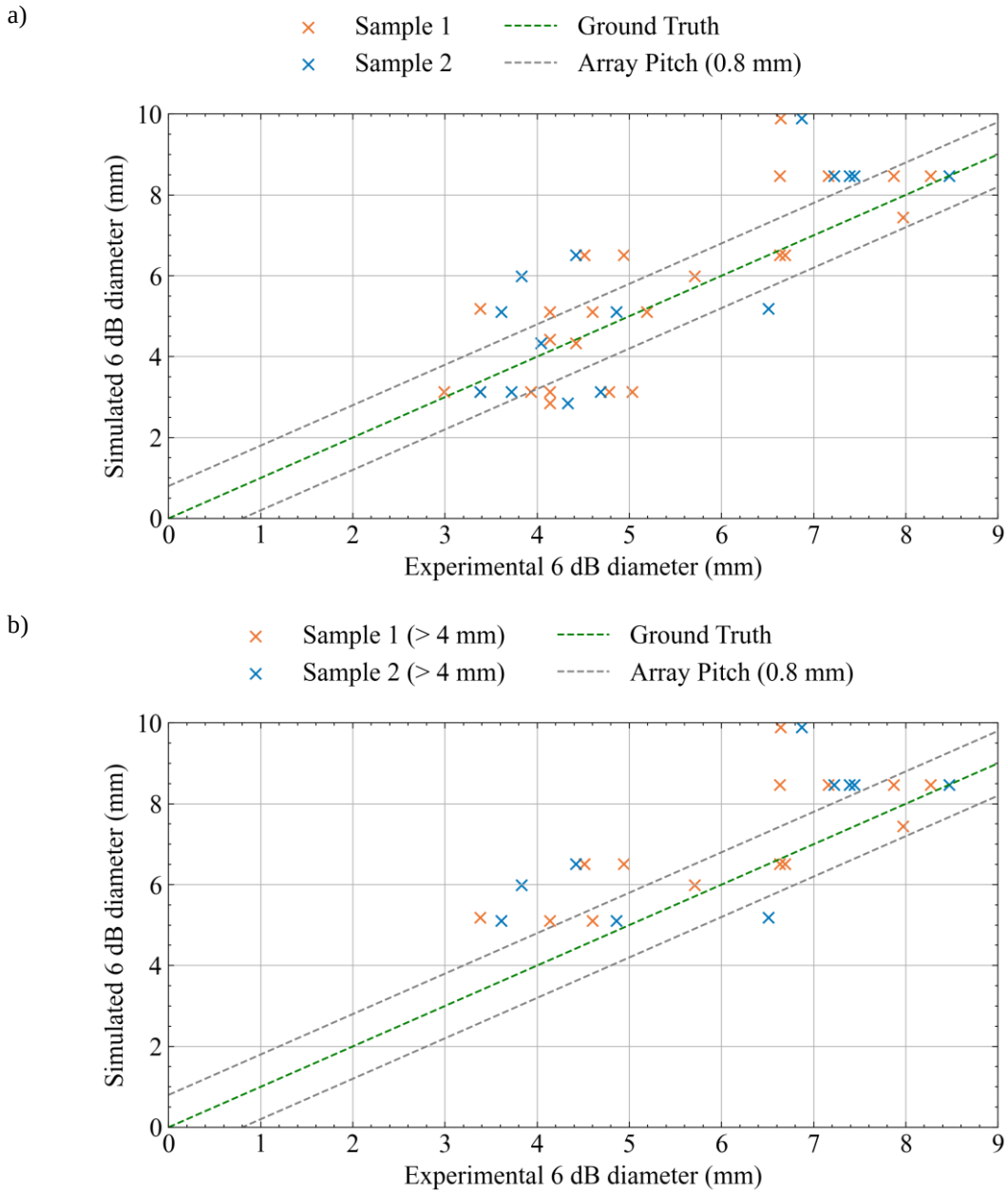


Figure 58: Comparison of sizing for synthetic and experimental data for all defect sizes (a) and defect diameters above 4 mm (b).

The comparison reveals that synthetic responses tend to yield larger defect sizes than experimental responses when employing the 6 dB drop method. Since our model's ground truth during training was based on synthetic response masks, it becomes apparent why there exists a propensity to overestimate defect sizes in our model; simulated responses tend to produce spatially larger defect responses than

experimental. This disparity can be attributed to a combination of factors, CFRP acting as an acoustic collimator is likely the most impactful cause. Due to the anisotropic nature of CFRP, attenuation increases significantly as the propagation angle to normal increases [178]. This effect was not captured in the semi-analytical simulations, as it did not account for the anisotropic acoustic attenuation. Consequently, these simulations resulted in increased acoustic beam spread and larger defect responses in the simulated domain compared to the experimental domain.

Table 21: Comparison in sizing of experimental and synthetic responses with the 6 dB drop method. Where R is the ratio of synthetic response to experimental response. $\overline{R}$ gives the mean average of R for a given defect diameter.

Sample 1					Sample 2				
Defect Diameter (mm)	**6 dB drop diameter**		$R = \frac{synth}{exp}$	$\overline{R}$	**Defect Diameter (mm)**	**6 dB drop diameter**		$R = \frac{synth}{exp}$	$\overline{R}$
	Experimental response	**Synthetic response**				**Experimental response**	**Synthetic response**		
9	6.63	8.47	1.28	1.21	9	8.47	8.47	1.00	1.18
	7.87	8.47	1.08			7.22	8.47	1.17	
	7.16	8.47	1.18			7.39	8.47	1.15	
	8.27	8.47	1.02			7.44	8.47	1.14	
	6.64	9.89	1.49			6.87	9.89	1.44	
7	4.51	6.51	1.44	1.22	6	6.51	5.19	0.80	1.26
	4.94	6.51	1.32			4.86	5.11	1.05	
	4.51	6.51	1.44			3.61	5.11	1.41	
	6.69	6.51	0.97			3.83	5.99	1.56	
	7.97	7.44	0.93			4.42	6.51	1.47	
6	3.38	5.19	1.53	1.18	3	3.38	3.13	0.93	0.83
	4.14	5.11	1.23			3.72	3.13	0.84	
	4.60	5.11	1.11			4.69	3.13	0.67	
	5.71	5.99	1.05			4.33	2.85	0.66	
	6.63	6.51	0.98			4.04	4.33	1.07	
4	4.78	3.13	0.65	0.82					
	3.93	3.13	0.80						
	5.03	3.13	0.62						
	4.14	4.42	1.07						
	5.19	5.11	0.98						
3	3.38	3.13	0.93	0.88					
	4.14	3.13	0.76						
	2.99	3.13	1.05						
	4.14	2.85	0.69						
	4.42	4.33	0.98						

To rectify this domain disparity, a constant *R* (given as the ratio between synthetic and experimental response diameters) can be computed based on the disparities in the 6 dB drop method for each defect, as detailed in Table 21. The correction factor is determined by the mean R calculated across defects with diameters exceeding 4 mm.

The resulting correction factor is 1.21 (as previously noted, defects with diameters ≤ 4 mm were deemed too small to be accurately sized with this experimental setup and the 6 dB drop method). When this correction factor is applied to the defect sizes determined from the model predictions, a significantly improved agreement with the known defect sizes is achieved, as shown in the Figure 59. This correction results in a reduction of the MAE by 58%, bringing it down to 0.88 mm. Consequently, defect sizing exceeds the accuracy of the 6 dB drop method significantly by 35%.

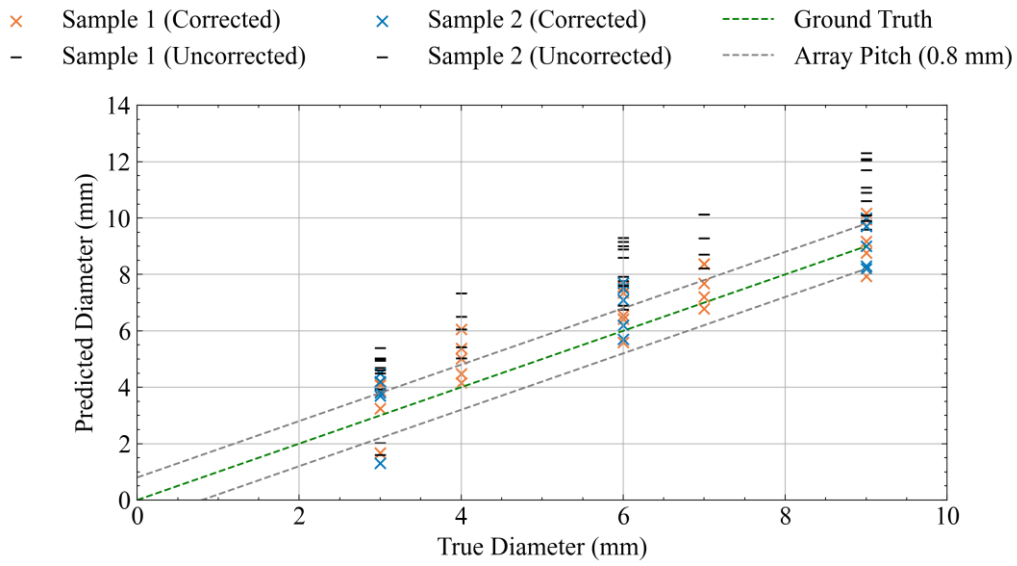


Figure 59: Corrected sizing results for the U-Net predictions.

5.5 Out-of-Distribution Testing

The model utilised in this study underwent supervised training, a process commonly employed in ML where the model learns patterns and relationships from labelled training data. Typically, supervised models excel in predicting examples that fall within the distribution of the training data. In this research, the training data was generated from simulated data, designed to replicate the geometry and characteristics of FBHs observed during testing. However, it's important to note that naturally occurring defects may exhibit significant variations in geometry. While defects in

composite materials often manifest in-plane, their characteristics can vary widely. Training a model to generalise across such diverse conditions would likely necessitate a substantially larger and more varied training dataset.

To give an insight into the model's generalisability to different defect types and geometries, its sizing and in-plane localisation performance was also evaluated on 15 square, 6 mm wide, Polytetrafluoroethylene (PTFE) inserts from a different sample. These different defect types can be considered a semantic distribution shift from the FBHs seen in training, and therefore categorised as “near Out-of-Distribution” [179]. 5 of the defects were located within 2 plies of the front wall surface, 5 were in the middle of the component and 5 were located within 2 plies of the back surface. This highlighted a limitation of the model, in that defects very near to front or backwall surfaces proved too challenging to segment. To account for this the front and back wall responses were removed. The results of this evaluation are detailed in Table 22, providing valuable insights into the model's performance for defects out of its training distribution.

Table 22: Sizing and in-plane localisation results for out of distribution test defects.

Width (mm)					Deviation in X (mm)	Deviation in Y (mm)	Absolute Displacement (mm)
True	Predicted	Predicted Error	Corrected	Corrected Error			
6	5.18	-0.82	4.28	-1.72	-0.18	-1.6	1.61
	6.69	0.69	5.53	-0.47	-0.24	-0.14	0.28
	8.98	2.98	7.42	1.42	0.08	0.47	0.48
	8.98	2.98	7.42	1.42	0.06	-0.45	0.45
	6.65	0.65	5.49	-0.51	-0.09	-0.18	0.20
	6.60	0.60	5.45	-0.55	-0.15	-0.21	0.26
	1.96	-4.04	1.62	-4.38	2.22	1.86	2.90
	7.11	1.11	5.88	-0.12	-0.44	0.07	0.45
	8.80	2.80	7.27	1.27	-0.03	0.51	0.51
	4.23	-1.77	3.50	-2.50	-0.42	-1.23	1.30
	5.99	-0.01	4.95	-1.05	-1.44	-0.69	1.60
	3.10	-2.90	2.56	-3.44	-0.82	-0.6	1.02
	8.80	2.80	7.27	1.27	0.09	0.1	0.13
	6.50	0.50	5.37	-0.63	-0.5	-0.07	0.50
	5.06	-0.94	4.18	-1.82	-1.04	-0.85	1.34
MAE		**1.71**		**1.50**			**0.86**

Whilst this does not serve to test the wide variety of defect geometries which occur naturally, the analysis provides insights into the model's efficacy to generalise to defects seen outside of the training distribution. A MAE of 1.50 mm for defect width was observed after the correction factor was applied. This represents a 71% increase in sizing error compared to defects within the training distribution. In-plane localisation performed much better with an MAE of 0.86 mm, a 51% increase in localisation error however, this is still in line with the element pitch. As these inserts were embedded pre-cure it is not possible to extract a true measured ground truth of through-thickness localisation as is the case for FBHs, this analysis has therefore been omitted. These results highlight a limitation of the proposed model and training regime. Synthetic data effectively addresses the challenge of acquiring labelled training data and can train an effective model for expected defects. However, for defects outside of the training distribution, there is a significant drop in performance. Furthermore, for edge cases such as defects near geometric features, additional pre-processing steps may be required; further limiting the generalisability of the model. Most DL work applied to NDE has a very specific and controlled application, and there's a notable challenge in finding models that have demonstrated effective generalisability across various materials, defect types, and component geometries. In future work the authors hope to further expand the synthetic training data to encompass a far wider range of defect types and geometries whilst also simulating different component geometries, accounting for edge cases such as near front and back wall responses. With a much larger synthetic training set it is hoped that a far more generalisable model can be trained.

5.6 Conclusion

This chapter proposes the use of a 3D U-Net to size and localise defects in CFRP by segmenting volumetric ultrasound data. Defect sizing is a crucial piece of information for evaluating defects against standards and acceptance criteria whilst accurate localisation is beneficial if re-work is required. A key benefit of the approach is that the use of volumetric ultrasound data allowed for through-thickness and in-plane localisation whilst removing the requirement for gating and reducing pre-processing. Such gating and amplitude threshold selection is often performed manually by the NDE operator leading to significant data errors if gates and thresholds are incorrectly set.

Simulations were used to generate synthetic data and ground truth segmentation masks for training. This was a key requirement and allowed for the training of a segmentation model in a fully supervised manor. Experimentally collected ultrasonic responses from manufactured reference defects were used for testing. Sizing and in-plane localisation were evaluated against the widely accepted 6 dB drop standard, and through thickness localisation was compared to the measured ground truth.

The models' depth-wise localisation showed excellent results with a MAE of 0.08 mm. In-plane localisation had good agreement with the accepted 6 dB drop standard with a MAE of 0.57 mm. The significant resolution differences in the spatial and temporal domains resulted in differences of error scales for in-plane and depth wise localisation. This is a limitation of using a fixed pitch array, but the errors in-plane are reasonable when compared to the array pitch of 0.8 mm, which is the limiting factor for spatial resolution.

The 6 dB drop consistently undersized experimental defects greater than 4 mm in diameter whilst the U-Net produced segmentation masks that consistently oversized defects. The U-Net's oversizing was consistent across defect sizes, showing that it was less reliant in maximum amplitude, making it more robust to sizing defects smaller than the width of the acoustic beam, which proved inconsistent for amplitude drop methods. In industrial settings other factors such as parallelism of defects to the inspection surface would also impact the maximum signal response, introducing further inaccuracies to amplitude-based sizing methods. Upon investigation of the synthetic and experimental data domains it was evident that the experimental responses gave rise to smaller defect responses. By correcting for this disparity between the source and target domain using a single correction factor it became possible to reduce the MAE for defect sizing from 2.09 mm to 0.88 mm. The corrected defect sizing from U-Net gave a 35% reduction in MAE sizing compared to the commonly accepted 6 dB drop method. Despite this the 6 dB drop method is based on physical understanding of defect responses and whilst it has limitations the results are directly explainable. Whilst the generation of a segmentation map aides in providing some explanation for defect sizing and is more interrogatable than a regression model, it still relies on a deep learning approach which is less explainable than an amplitude-based threshold. Even so, a clear benefit of the volumetric segmentation is that it can be translated directly into a computer-aided design file which could dramatically increase the efficiency of subsequent report generation and simulation-based testing of components.

While the study demonstrated promising results in defect sizing and localisation, it's essential to acknowledge a current limitation: the absence of real defects from

industrial manufacturing processes for testing. Although out-of-distribution cases were examined using PTFE inserts, evaluating performance on naturally occurring defects would be advantageous. The out-of-distribution testing underscored the necessity for a more extensive distribution of training data. Future work aims to address this by expanding the dataset, incorporating different probe frequencies, and testing on naturally occurring defects with irregular shapes. Furthermore, the objective is to integrate this research with previous work on defect detection to develop an end-to-end system for automated NDE data processing in industrial manufacturing environments.

6 Self-Supervised Learning Segmentation

6.1 Introduction

While the 6 dB drop method remains a cornerstone in defect sizing, advancements in DL techniques offer promising avenues for enhancing defect localisation and sizing accuracy. By leveraging the capabilities of fully supervised volumetric methods, the previous chapter has made significant strides in addressing the complexities associated with defect characterisation and quantification. Analysing volumetric data offers a wealth of information for defect characterisation that surpasses what can be achieved through individual image analysis alone. Additionally, this approach proves advantageous in reducing manual preprocessing tasks, such as gating out structural responses, which are often labour-intensive. Moreover, volumetric segmentation masks of components open avenues for various downstream tasks, such as building digital twins [166], performing FEA, and alleviating the burden of report generation.

However, the previous fully supervised method faces limitations due to the necessity of hard to acquire labelled training data, a common requirement for any fully supervised training approach. In many NDE applications, obtaining accurately annotated labelled datasets of real defects is challenging and often not possible, and is one of the main barriers for applying DL to NDE [14]. To address this scarcity issue, work in previous chapters utilised synthetic training data generated from simulations. By controlling simulation parameters, accurate labelling of segmentation ground truth could be achieved automatically. However, accurately simulating the full distribution of defects and their variations is computationally demanding, and it is challenging to ensure fidelity between the simulated and experimental domains [180]. As a result, if

defect segmentation could be performed without the reliance on large positively labelled datasets, it would offer significant advantages.

Self-Supervised Learning (SSL) [181], [182] is a method for training DL models in a supervised manner without the need for labelled training data. Generally, labels are generated through auxiliary tasks or by leveraging inherent structures in the data itself, enabling the model to learn meaningful representations without explicit human annotation. By training in a supervised manner, models are often able to learn more detailed feature representations than unsupervised approaches. SSL introduces the ability to leverage large amounts of unlabelled data, reducing the need for costly and time-consuming annotation. Its versatility has meant that it has been applied broadly from computer vision to natural language processing and has demonstrated impressive performance in many notable DL tasks, such as with large language models [183], [184] and large vision models [185], [186]. In general, SSL works by formulating pretext tasks that require the model to predict certain aspects of the input data based solely on the input itself. These pretext tasks are designed to be easily computable from the raw data without the need for external annotations. By solving these pretext tasks, the model learns to extract meaningful features and representations from the data, which can then be transferred to downstream tasks. For example, in natural language processing, the model may be tasked with predicting missing words in a sentence (e.g., masked language modelling [187]) or predicting the next word in a sequence (e.g., language modelling [183], [184]). Similarly, in computer vision, SSL tasks may involve predicting the rotation, colorization, or spatial arrangement of patches within an image [188], [189], [190], [191]. Another prevalent strategy in SSL is contrastive learning, where the model learns to differentiate between positive and negative pairs

of data samples. By maximising the similarity between positive pairs (e.g., different augmentations of the same image) while minimising the similarity between negative pairs (e.g., augmentations of different images), such as Siamese networks [192].

This chapter presents a novel approach aimed at advancing the detection, localisation, and segmentation of defects within ultrasonic volumes. The approach leverages the capabilities of SSL coupled with a 1D CNN to achieve 3D segmentation of defects from volumetric ultrasonic testing data of composite components. We employ pretext learning to predict distributions of amplitudes from ultrasonic series. During the inference stage, the pre-trained 1D CNN is deployed to flag any regions within the component that exhibit anomalous behaviour. This process capitalises on the insights gleaned from the pretext learning task, where the model familiarises itself with clean samples. Unlike traditional approaches, which often necessitate extensive positive training examples, this methodology operates on the principle of anomaly detection rather than classification into specific defect classes. By reframing the problem as one of anomaly detection, rather than attempting to categorise defects into predefined classes, we circumvent the significant challenge of acquiring an extensive set of positive training examples. Moreover, this approach offers the advantage of being defect-agnostic, thereby mitigating concerns regarding the generalisability of the model to novel defects. This characteristic alleviates the need for meticulous fine-tuning and ensures that the model remains robust across a range of defect types, as long as they produce an anomalous response, eliminating the burden of adapting the system for each new defect type encountered. Overall, by adopting this innovative methodology, critical limitations in the application of DL to NDE are addressed, paving the way for more efficient and robust defect detection in ultrasonic inspection

processes whilst giving information on not just detection but also complete volumetric localisation and segmentation which has only been previously achieved with fully supervised training [193].

Section 6.2 of this chapter outlines the data used and any pre-processing requirements. In Section 6.3, the pretext learning is detailed. The results and discussion are presented in Section 6.4.2. This, for the first time, introduces an SSL method for volumetric defect segmentation of ultrasonic testing data, offering several advantages:

- No defective training data is required.
- Minimal preprocessing.
- Geometric features are retained.
- Enhanced generalisability for defect detection is achieved by reframing the problem as anomaly detection.

6.2 Data

Additional CFRP samples of varying thicknesses, supplied by Spirit AeroSystems, were employed in this study. For pretext learning, samples verified as defect-free through ultrasonic inspection, analysed by an NDE operator, were selected. These samples were segregated into distinct datasets for training, validation, and testing purposes. During the inference phase, the previously introduced defective samples with FBHs were used for testing (see section 2.2). Additionally, as used for the out-of-distribution test for the supervised segmentation method, a final stepped sample with square 6 mm wide PTFE inserts presents a more challenging geometry and defect responses, offering a representation closer to naturally occurring defects. A summary of the samples and dataset characteristics is presented in Table 23.

Data acquisition was conducted using an automated ultrasonic system centred around a 64 element, 5MHz Olympus linear phased array roller probe, with an element pitch of 0.8 mm. Further details regarding the experimental setup and composite samples can be found in section 2.2. The segmentation methodology employed in this study was intentionally designed to be versatile, thereby maximising its applicability across different scenarios. Consequently, only minimal generic data preprocessing was applied. The data preprocessing involved enveloping the signal using the Hilbert Transform (see section 2.3). Notably, techniques such as TCG, peak-alignment, and gating out of geometric features were deliberately omitted. This helps to ensure more consistent and reproducible results across different experimental setups, as it does not require adjusting for specific setups or components, leading to less calibration and setup time. It also reduces the risk of human error and biases that can be introduced during these (potentially) manual processes. Consequently, this leads to more robust and reliable data analysis.

Table 23: Summary of samples used.

	Sample	Thickness (mm)	Dataset Size [Probe | Time | Frames]	Details
Pretext Learning	Clean 1	2.75	122|350|260	Training
	Clean 2	4.25	122|450|260	
	Clean 3	4.25	122|450|260	
	Clean 4	6.00	122|600|260	
	Clean 5	6.00	122|600|260	Validation
	Clean 6	8.60	122|700|260	Test
Inference	Defective 1	8.60	183|700|260	15 FBH
	Defective 2	8.60	305|700|230	25 FBH
	Defective 3	7.50,9.60,11.80	488|1050|112	15 Inserts

6.3 Pretext Learning

Typically, ultrasonic signals are considered as A-scans, which is often appropriate. However, when inspecting large components C-scan images often provide the most impactful and clear information about the structure of the component. C-scans compress information in the time-domain either through amplitude imaging or amplitude indexing (Time-of-Flight). This loss in temporal information is done as a result of trying to maximise comparative spatial information, which can often be the most helpful when evaluating components of known geometry, particularly for composite components where structural noise can show significant variability in the time-trace. Operators will often produce C-scan images at different thickness gates to compare different through-thickness slices. Intuitively therefore it can be deduced that considering comparative spatial information for given time-windows is as, if not more, important than direct temporal comparisons along a single time trace. Ideally spatial comparisons would be conducted at every depth but for manual interpretation this is intractable due to the substantial amount of data produced.

Mechanised linear phased array scanning is commonly employed in the inspection of large-scale industrial components [7], [8]. Arrays operate within acceptable tolerances of element sensitivity. Whilst compensation can be done to account for inter-element variations within arrays through calibration, some level of variation between elements is likely to remain during scanning. Considering this and the importance of spatial comparison discussed earlier, it follows that it is appropriate to analyse a singular series through a component at a particular time step (Figure 60), and is the basis for this work. This approach minimises variations from array elements by concentrating

on the same reception element or sub-aperture, while facilitating spatial comparison at a given depth. Furthermore, this methodology can be easily conceptualised as a self-supervised learning task. By treating it as a 1D task, we not only reduce model size and complexity but also benefit from a substantial increase in training data abundance compared to using 2D or 3D training sets from the same sample availability.

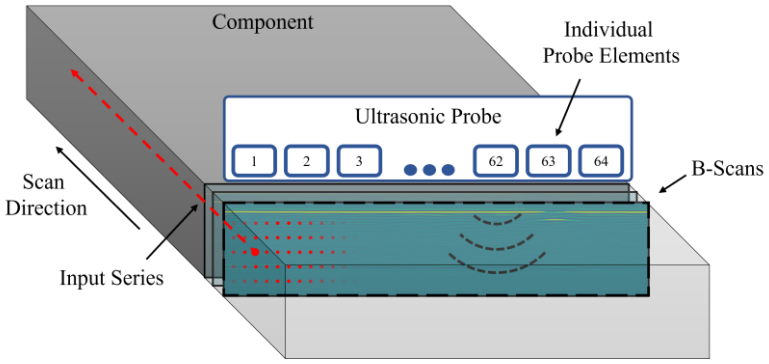


Figure 60: Diagram illustrating an example of a through-component series, as examined for the pretext learning task.

For the pretext training task in this study, inspiration is drawn from language modelling (where a model tries to predict the next token in a sequence), aiming to predict the next value in the series for a clean sample. While alternative methods such as contrastive learning or generative learning could be utilised, the task of predicting the next value lends itself well to in-process inference and, more broadly, to the downstream inference task of volumetric segmentation.

Preliminary trials, focused on single-value prediction, which demonstrated good performance on the pretext task but performed poorly on the inference task due to the lack of accountability for variability within sequences. This made establishing appropriate thresholds challenging and often lead to noisy inference results. For instance, while high-amplitude areas of the scan (such as the front wall) might yield

relatively accurate predictions, the significant variability posed an issue. Although variable or percentage-based thresholds provided some benefit, the underlying problem persisted. To enhance this method, a more sophisticated Probabilistic Neural Network (PNN) approach was employed. In this approach, the model attempts to predict the distribution that corresponds to the likelihood of the expected next value in the scan sequence, as depicted in Figure 61. The distribution shows the expected probabilities of different values based on the model's learnt knowledge and the input series, where the mode represents the most likely expected value. This allows the model to account for areas of prior variability or lack of variability by widening or tightening the distribution. Thresholds can then be set as confidence intervals against these distributions, automatically accommodating series variability. We employed a sequence length of 64 values for prediction, aiming to strike a balance between having a sufficiently large receptive field to learn about the distribution and patterns of amplitude responses within the component, while also ensuring that the sequence length is not excessively long, which could limit inference to large components and minimise access to training data.

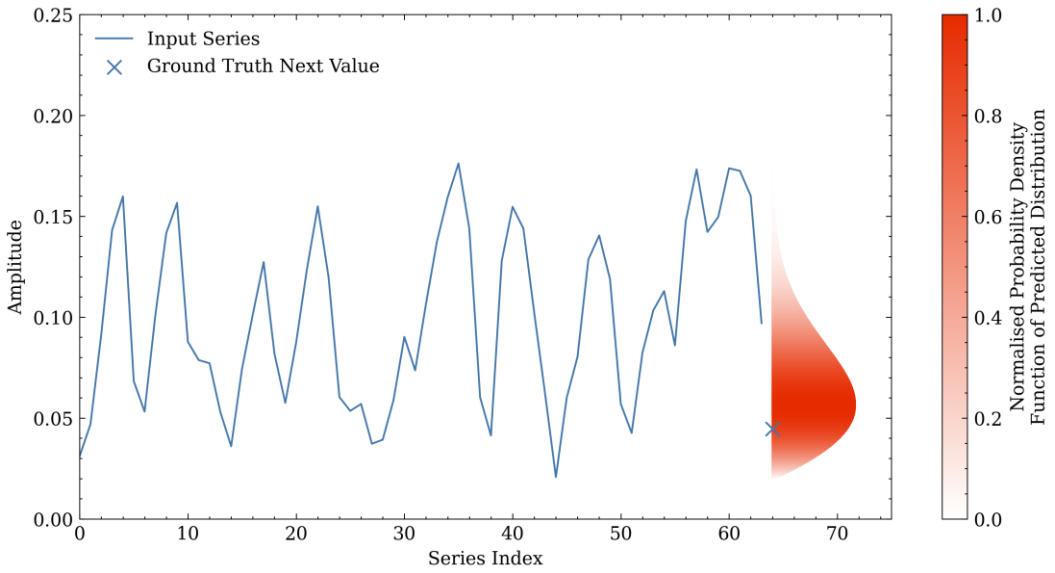


Figure 61: Example of the models' predicted distribution for the next value from a given input series and comparison to the true value.

Various models for series forecasting exist, ranging from DL methods such as Recurrent Neural Networks (RNNs) and Transformers to probabilistic approaches like Gaussian Processes [194], which can be adapted for probabilistic prediction based on specific requirements. In this study, a 1D multi-head CNN, inspired by InceptionTime [195], was adopted for the architecture, with short series (<128) shallow networks with a lower number of shorter filter lengths being deemed sufficient. This has enabled the 1D CNN to be a lightweight model, with 486,242 trainable parameters occupying just 1.94 MB of memory.

The model probabilistic approach is characterised by predicting the scale and concentration parameters of a two-parameter Weibull distribution as outputs. The choice of a two-parameter Weibull distribution enables the modelling of various distribution shapes for continuous positive values (as consistent with enveloped amplitude values), accommodating non-symmetric distributions through both left and right skewed data. While this model and method proved effective for the downstream

segmentation task (please see section 6.4.2), the exploration and evaluation of alternative probabilistic regression methods and architectures which could lead to the prediction of better fitting distributions has been left for future work. The model architecture is depicted in Figure 62, where each convolutional block consists of a convolutional layer with the specified kernel size and stride of 1, followed by a convolutional down-sampling layer with a stride and kernel size of 2. A LeakyReLU activation function is employed between each convolutional and fully connected layer.

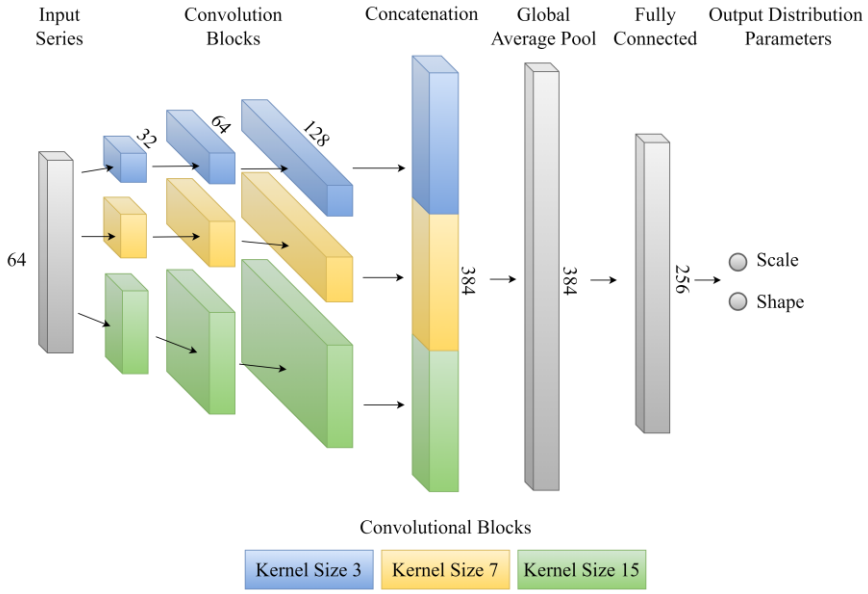


Figure 62: Probabilistic CNN architecture.

During pretext learning, clean samples 1-4 were used for training with a batch size of 65536, whilst clean samples 5 and 6 were kept as holdouts for the validation and test sets respectively (Table 23). Adam optimiser [70], with a learning rate of 1×10^{-6} is used to minimise the Negative Log-Likelihood (NLL) loss for the Weibull distribution, as given by equation (17), where $f(a, b \mid x_i)$ is the Weibull probability density function parameterised by Scale (a) and Shape (b). To minimise overfitting, a patience of 3 epochs was used when evaluating the validation set to determine early stopping.

$$Weibull\ NLL\ Loss = -\log \prod_{i=1} f(a, b \mid x_i) = -\sum_{i=0}^{n} \log f(a, b \mid x_i) \quad (17)$$

During training the data is down sampled in the time domain by every 5 samples (approximately 30 µm in depth within the CFRP samples). This was done as points next to each other in the time domain exhibit minimal variation because of the high ultrasonic sampling rate; offering limited additional information to be learnt and an increased computational cost during training.

During training, a hyperparameter which arises for this problem is the stride for sampling data during training. Consider a single full-length scan series: the rate at which this full length is sampled for new training samples is determined by the stride of the window applied, as demonstrated in Figure 63. In DL, it is generally advantageous to maximise the amount of training data available. Using a stride of 1 achieves this by providing maximum available training data. However, whilst each sample is different by a single point, there exists significant overlap between neighbouring series, exposing the model to very similar data during training. In such cases, the model may not gain additional information from the additional samples and could become prone to overfitting the training data.

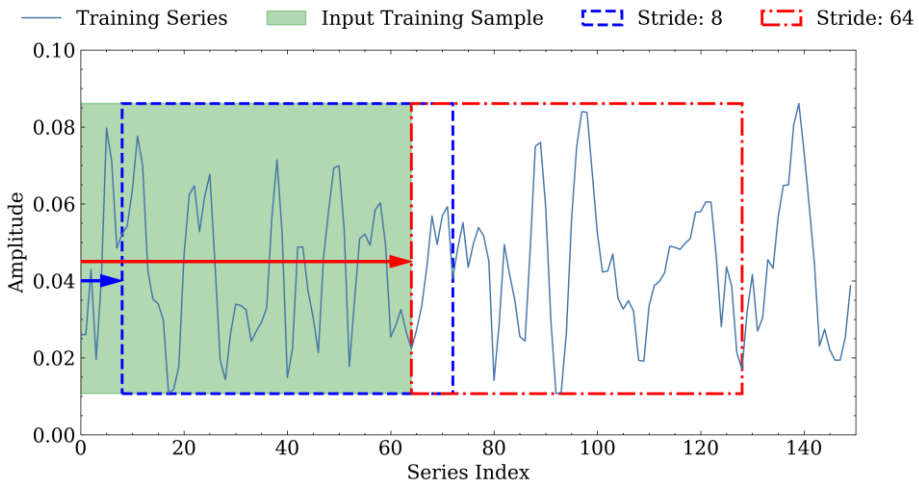


Figure 63: Demonstration of the impact of stride (8 and 64) when sampling the training data.

Alternatively, a stride length of 64 can be employed, where each training example represents a distinct series without information sharing due to overlap with other training examples. While this reduces overlap, it also significantly reduces the available training data. Picking an appropriate stride for sampling training series is therefore a trade-off between training set size and overlap in the training data, potentially leading to overfitting. To explore this relationship, we tested the following stride values: 1, 2, 4, 8, 16, 32, 64, 128, 256 which corresponded to dataset sizes of 74.12 M, 37.06 M, 18.53 M, 9.46 M, 4.92 M, 2.65 M, 1.51 M, 0.76 M, 0.38 M respectively. It is worth highlighting that an advantage of using series in this way is the production of very large training datasets compared to typical image-based ML research in NDE. For this test, the stride of the validation set was fixed at 64, and the testing stride was set to 1 to maximise the test set size (which also matches the requirement during inference). The model was trained three times for each stride. For each training iteration, the mean log-likelihood was recorded across the test set, and the mean and standard deviation of the results from the three training runs are reported in Figure 64. A higher log-likelihood between the predicted distribution of the

expected value and the ground-truth measured value indicates that the model's prediction is better aligned with the measured value. This suggests that the model is better at capturing the underlying patterns or relationships in the data, leading to more accurate and precise predictions.

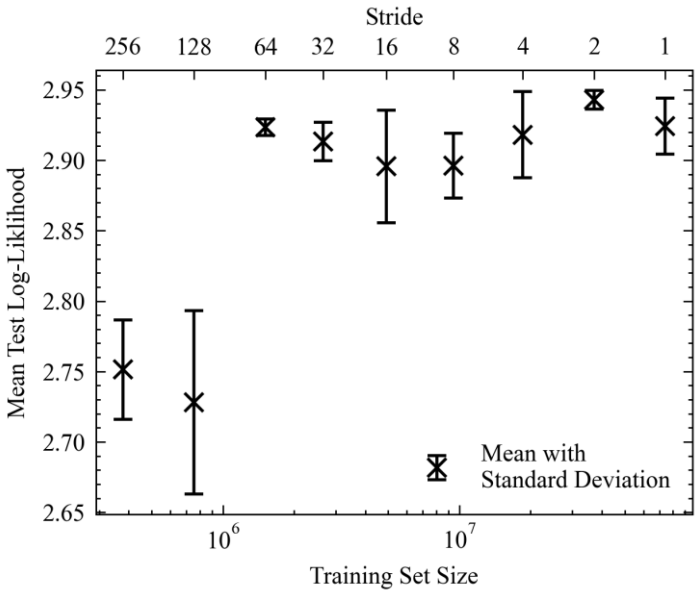


Figure 64: Test set mean Log-Likelihood for varying sampling strides.

As depicted Figure 64, a stepped improvement is observed when the training datasets exceed 1 million samples (stride < 128). However, beyond this threshold, the model appears to exhibit diminishing returns from additional training data. Similar trends to this have been documented in the existing literature [196]. Perhaps surprisingly, there is not a consistent increase in performance for a reduction in stride. Whilst this could indicate model saturation, there could also be a detailed relationship between training data and overfitting at play. Notably, strides of 64 and 2 exhibited lower deviations in results, coupled with good performance, suggesting the emergence of stable solutions. It's important to note that different pretext learning models and training datasets may exhibit varying relationships with stride length. Nonetheless, this analysis underscores

the significance of tuning this parameter and provides insight into the required training dataset sizes. In future work, it is hoped to explore this further and assess its impact on inference performance.

6.4 Inference

6.4.1 Methodology

By training on defect free data, the pretext learning phase yields a model which predicts the expected distributions of the next defect free value for a given time index and reception element. By leveraging this pre-learnt information, the 1D model is utilised by applying it to each preceding 64 series of volumetric data to forecast the distribution for each point in the subsequent B-scan. This enables anticipation of the distribution of data points across the entire next B-scan under the assumption of a defect-free component.

During inference each volumetric frame (B-scan) is processed in a sequential manner, comparing the predicted clean B-scan to the measured experimental result at every point in time for each receptive element or sub-aperture of elements. Since the model is trained only on clean data to predict the next clean value in the series, if there is agreement between the values in the predicted frame and the measured experimental frame they are marked as defect free. The inference window is then moved to the next frame in the data and the process is continued. This acts to update the predictive model with the most current information and keeps good alignment with variations seen in specific samples/scans.

However, if the next B-scan contains defective voxels this will result in an increased amplitude response around the defect which is outside the distribution of clean sequences seen in the prior learning. As a result, when comparing the predicted (clean) frame to the measured (defective) frame there will be poor agreement locally around the defect response. These voxels can therefore be marked as defective within the volume; locally segmenting the defect. For the series used for the subsequent frame, prior predictions not marked as defective are treated as normal with the series window sliding along from the experimental series. However, for defective voxels, the mean is taken from the probabilistic output as an a priori estimate of a clean response and is used when evaluating future sequences. Doing this, ensures that the model is always predicting expected distributions of clean responses based on prior clean sequences or the closest estimate to clean values, and errors as a result of defective responses do not impact future predictions, as consistent with the pretext learning stage. This process is outlined in Figure 65.

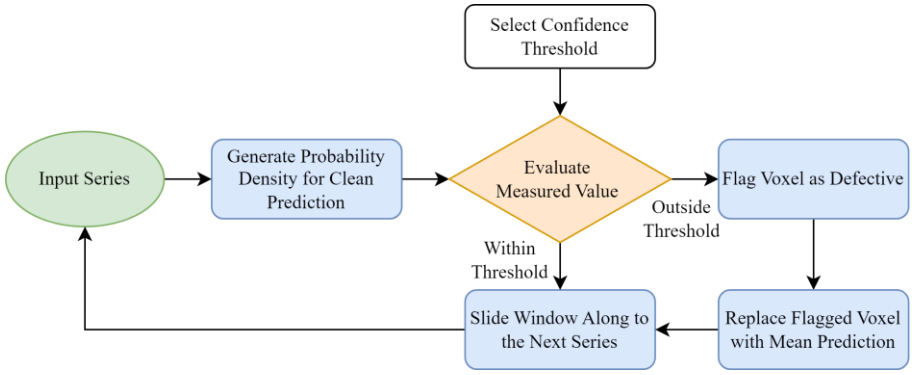


Figure 65: Flowchart of sequence prediction for inference.

Given the probabilistic output, a threshold must be set to evaluate the model's clean prediction against the measured value to determine if a voxel is defective. This threshold can be set as a confidence threshold on the predicted clean output

expectation, making it adaptive to variations in series amplitude responses. There are different ways to determine an appropriate confidence threshold such as experimental calibration. An alternative is to base it off an allowable false-call rate. Since this detection analysis is for the entire volume and due to the resolution difference in the spatial and time domain, the number of voxels is far larger (700 times for test samples 1 and 2) than for image level analysis. For the same expected number of absolute false calls for image analysis the false call rate for the volume would therefore need to be far lower (due to the increased number of voxels). To account for this a confidence threshold is chosen based on a much lower allowable false-call rate than would be expected for image analysis. It is important to note that this false-call rate is done on a per voxel basis and not per defect basis. In this study results are presented for false-call rates ranging from 1% to 0.00001%. Even at the aggressive lower false-call rate of 0.00001, sample 2 which has 5.6 M voxels would still be expected to produce approximately 6 false-calls. Whilst for these samples this is a low number, for larger parts this would scale cubically for any increase in the number of B-scans as a result of a longer scan.

For volumetric inspection, the inference process can be completed just following the forward scan (as most applicable to in-process inspection), herein denoted as the forward sweep, or post scan the analysis can be completed in reverse – simulating the scan from the other direction, herein denoted as the backwards sweep. Taking the logical AND of both passes as the final segmented volume acts to increase confidence in defective predictions by having them detected twice. This helps to further remove false positive indications and cleans up the segmented response as seen in Figure 67.

To further mitigate the impact of false-calls, a minimum area threshold is applied to detect only defects above a critical size and limit low-area voxel clusters resulting from the probabilistic nature of predictions. For CFRP applications, defects generally align parallel to the ply orientation. It is this area which is typically used to assess critical defect size. Area thresholding is calculated using equation (18), where *Area Opening* refers to the *skimage.morphology.area_opening* [197] which, removes all connected components smaller than the *Filter* for the volume. The connectivity determines what neighbours are considered connected components, for this application all 8 neighbouring voxels are considered for a 2D slice through the volume. For other applications and materials, where the defects are not primarily in-plane, 3D connected components may be more appropriate.

$$Thresholded\ Volume = Area\ Opening(Volume[:, depth, :], Filter, Connectivity) \quad (18)$$

The minimum defect size threshold can be tailored based on the specific application requirements. For testing purposes, where the minimum defect size was 3.0 mm in diameter, the area threshold was adjusted to exclude any indications smaller than this. For applications where thin crack-type defects are typical, such as in metals, it may be beneficial to consider a more complex shape-based thresholding method, such as evaluating the defect aspect ratio to account for extreme geometries. However, for composites, which generally exhibit in-plane defects, area thresholding is deemed appropriate and is used by Spirit AeroSystems to determine defect criticality. Area segmentation is completed as the final processing stage. The complete workflow for volumetric instance segmentation is illustrated in Figure 66. Details on the impact of each processing step for different thresholds can be found the in the following Results section (Figure 69) and Table 28 of the Appendix.

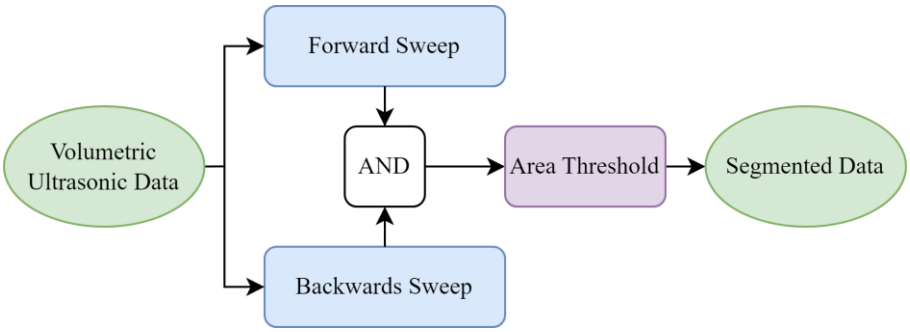


Figure 66: Flowchart of the methodology overview for complete volumetric segmentation.

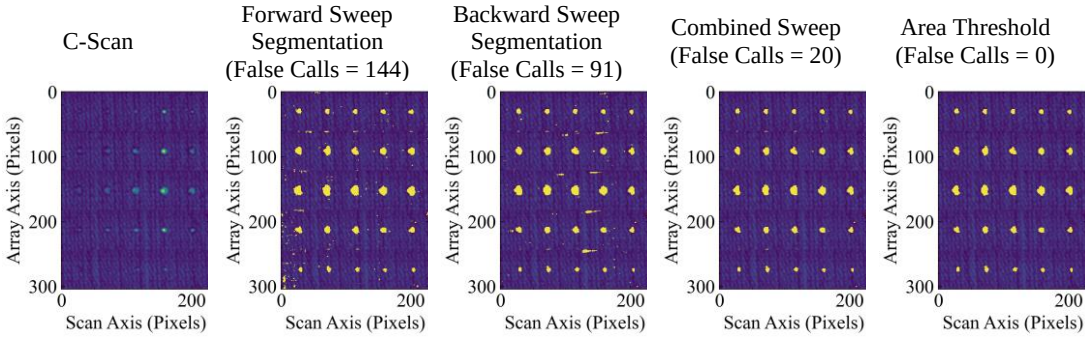


Figure 67: Demonstration of the impacts of different post-processing steps for defect sample 2 and threshold of 0.9999999.

To process the first frames where there are not 64 prior frames for prediction, padding is used to populate the remaining missing information for each sequence. Whilst zero padding, constant edge padding or reflect padding can all work as the combined sweeps clear up any errors as a result of not having enough prior information for predictions. They can produce artefacts for a single sweep at the edge of the scan due to a lack of prior information for the model to give an accurate prediction. Reflect and edge padding give the best results, reducing the reliance on sweeps to clean up edge artefacts. Figure 68 shows an example of edge artefacts from using zero padding compared to edge padding.

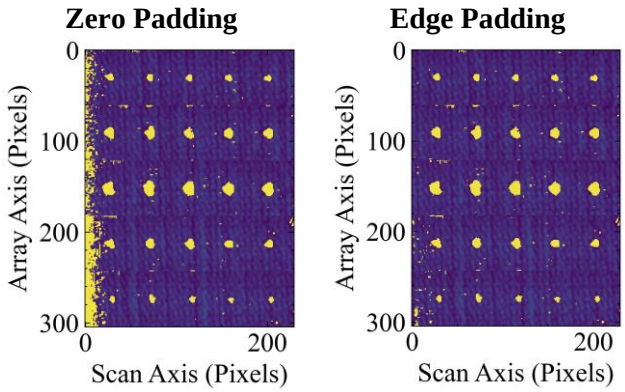


Figure 68: Example of edge artefacts for a single sweep when using zero padding compared to edge padding (sample: defect 2, threshold: 0.9999999).

Different defects exhibit varying characteristics. This method reframes the problem of positive defect identification as one of anomaly prediction, which is beneficial for generalisability to a wide range of defects. However, as the model has not learnt specific positive features of defect responses, any change in amplitude response, whether from a defect or geometric feature, can lead to an anomalous prediction. Real components often feature geometric elements such as step changes in thickness. While inference can still be conducted in such cases, analysis must be conducted in parallel with any geometric alterations so that they are present in the series used for prediction. Therefore, if there is a part with very complex geometry it may have to be sectioned for inference. To evaluate the method's performance under these conditions and to assess its effectiveness with different defect types, we applied the approach to a stepped sample with PTFE inserts (defective sample 3). These defects are inserted pre-cure and are more representative of naturally occurring defects compared to FBHs.

For inference the model used was the best performing during pretext learning. Inference is a sequential process, but the computational cost can be easily minimised by batching the predictions for a complete frame; due to the lightweight 1D CNN. To reduce the computational cost during inference, the time domain was down sampled

by a factor of 10 for testing. Previous work demonstrated that it is still possible to get very high levels of depth-wise resolution with this sampling [193]. Inference for a single frame took approximately 0.05s when batched. To complete the full processing pipeline of forwards and backwards sweeps and area thresholding took less than 35s for each sample. Testing was conducted on a workstation equipped with a NVIDIA GeForce RTX 3090 running the Pytorch [146] framework. The following section presents results for defect detection, sizing, localisation, and visualisations for qualitative examples of volumetric segmentation.

6.4.2 Results

6.4.2.1 Detection

Detection performance of the model was evaluated for thresholds ranging from 0.99 to 0.9999999, with results reported for each processing stage. Evaluated was conducted on amplitude C-scans of the segmented volumes to avoid the impact of repeat echoes stemming from defect indications. The ground truth defect mask was generated using manual identification of defects, with the 6 dB drop used for locating the centroid of each defect. The true defect sizes were then used around the centroids to mark out the defect areas. A defect is considered detected if the predicted mask overlaps with the ground truth mask. If there is no overlap the prediction is considered a false-positive. Across all thresholds and processing stages, all defects were successfully identified, with no missed detections; however, a notable decline in accuracy is observed as a result false positive indication. Whilst the absolute number of false positives can be large, they are often very small in size (as demonstrated in Figure 67) and can therefore be effectively removed with area thresholding, which can

be derived from critical defect size (as dictated by industrial testing standards). The results for detection accuracy are presented in Figure 69 (with a full breakdown available in Table 28 in the appendix). It is feasible to completely minimise false positives, thus achieving a detection accuracy rate of 100%, using the processing steps suggested in Figure 66 and a threshold of 0.9999999.

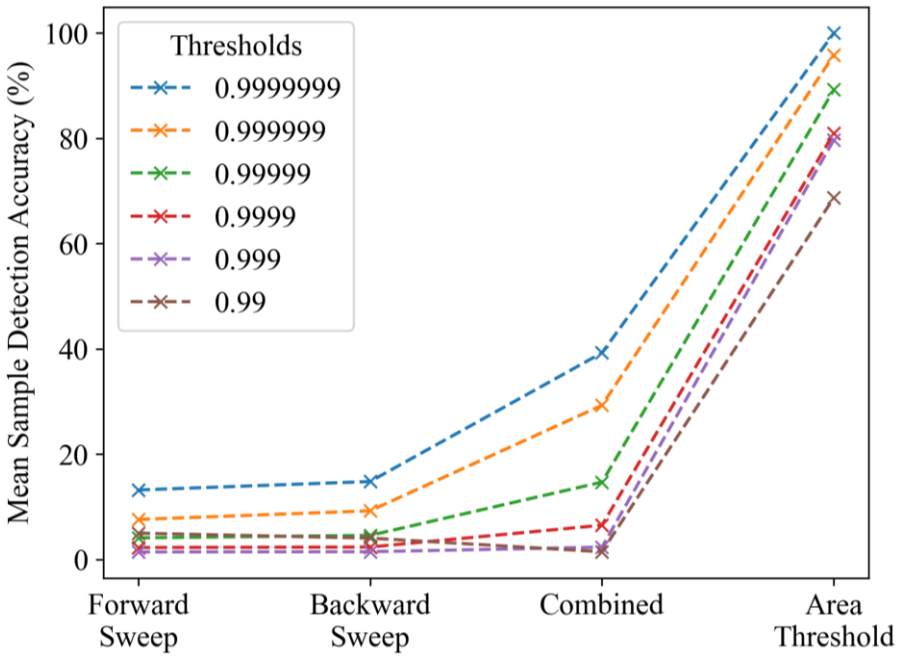


Figure 69: Defect detection accuracy for each threshold and processing step.

6.4.2.2 Defect Sizing

The impact of defect sizing has been assessed across the same range thresholds as used for detection, with the results summarised in Table 23. While the deviation of absolute error remains relatively stable, there exists a strong negative correlation between increasing the confidence threshold and a reduction in MAE for defect sizing. Notably, at a threshold of 0.9999999, the MAE of 1.41 mm, aligns closely with the findings of a previous study which reported errors in sizing inaccuracy for the 6dB drop method of 1.35 mm [193]. Moreover, this performance surpasses that of the fully supervised

3D U-Net method previously presented, prior to adjustments for differences between synthetic and experimental domains.

The method has been shown to only oversize defects. This stems from the underlying mechanism of detecting anomalous voxels, which does not always correspond perfectly in a one-to-one manner to actual defect size; due to impacts of ultrasonic imaging such as beam spread etc. The impact of this is visually demonstrated through an example B-scan in Figure 70. Consistency in oversizing defects yields two main advantages:

Safety Considerations: Oversizing defects, while potentially sacrificing some precision, inherently reduces the risk of overlooking critical flaws. In safety-critical industries such as aerospace or structural engineering, the consequences of under sizing defects can be severe, leading to structural failures or operational hazards. By consistently erring on the side of caution and oversizing defects, the method provides a safety buffer, ensuring that potential weaknesses are identified and addressed before they escalate into safety incidents.

Calibration Potential: The method's consistent error pattern facilitates calibration procedures, akin to those commonly employed in traditional amplitude drop sizing methods. Calibration involves establishing a correlation between measured responses and actual defect sizes, thereby refining the method's accuracy and reliability. The method's predictable oversizing behaviour lends itself well to calibration processes, allowing for adjustments that compensate for systematic errors. For instance, for a threshold of 0.9999999, when utilising defect sample 1 as a calibration reference, the MAE for Defect Samples 2 and 3 decreases to 0.58 mm, a 57% reduction in MAE.

Table 24: MAE for different thresholds.

Sample	Defect Width (mm)	MAE for given threshold					
		0.99	0.999	0.9999	0.99999	0.999999	0.9999999
Defective 1	9	2.75	2.25	1.41	1.27	1.13	1.02
	6	3.39	2.55	2.05	1.67	1.44	1.27
	3	3.78	2.91	2.60	2.21	1.83	1.75
Defective 2	9	3.14	2.38	1.86	1.49	1.20	0.93
	7	3.26	2.61	2.19	1.83	1.50	1.23
	6	3.18	2.62	2.10	1.74	1.37	1.23
	4	3.67	2.97	2.45	2.17	1.93	1.79
	3	3.93	3.39	2.94	2.58	2.32	2.06
Defective 3	6	3.06	2.32	1.73	1.46	1.29	1.21
Mean		**3.37**	**2.71**	**2.20**	**1.87**	**1.59**	**1.41**
Standard Deviation		0.66	0.61	0.69	0.67	0.66	0.68

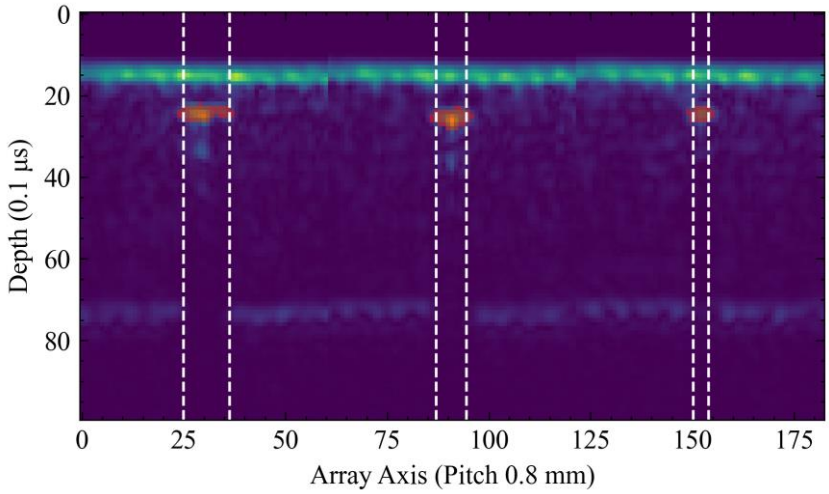


Figure 70: Example B-scan across multiple raster passes showing the voxels highlighted as defective and the corresponding true defect size in white. This is shown for defect sample 1 with a threshold of 0.9999999.

6.4.2.3 Localisation

For linear scanning of composite materials, achieving accurate volumetric localisation necessitates an understanding of both in-plane and depth-wise localisations. For depth-wise localisation, the evaluation focuses on FBHs, while excluding inserts. Inserts, positioned between layers pre-cure, pose a challenge in obtaining true ground truth depth measurements post-curing, unlike FBHs which can be accurately measured post

cure to obtain a ground truth depth measurement. Through-thickness measurements are derived from the mean thickness through the centroid of the segmented defect.

In contrast, acquiring a ground truth measurement for in-plane localisation presents its own challenges. Cumulative positional errors inherent in the experimental setup render the attainment of accurate ground truth measurements (<1.0 mm) infeasible. Consequently, this research adopts the 6 dB drop criterion as a reference standard for validating the agreement between segmented in-plane localisation and the 6 dB drop criterion. By doing this the in-plane localisation can be compared against an industry standard method as a validation benchmark. To assess in-plane localisation, the Euclidean distance between the centroid of the 2D projected segmentation mask and the 6 dB drop masks is computed.

The integration of both through-thickness and in-plane localisation provides a comprehensive understanding of the method's ability to locate defects within the volume. Table 25 presents the mean and standard deviation of results. The MAE for depth localisation surpassed that of in-plane localisation, with comparable standard deviations for both. Although this method showed improvement for in-plane localisation compared to previous fully supervised volumetric segmentation methods, it fell short of achieving the high accuracy levels seen in through-thickness depth localisation. This discrepancy might be partly attributed to the additional pre-processing steps used for the fully supervised method, such as peak alignment. Enhancing through-thickness accuracy could potentially be achieved by increasing the temporal sampling rate during inference, although for most applications, the current level of accuracy is likely sufficient. Both metrics demonstrate considerable

localisation performance, with MAEs well below 0.5 mm. This level of precision is likely more than sufficient for typical industrial rework scenarios, as it aligns with the accuracy required for precise tooling operations Additionally, the in-plane MAE is far below the element pitch, which is the limiting factor for in-plane imaging resolution. All in all, the model performs well in volumetric defect localisation.

Table 25: Localisation results.

Sample	Defect Width (mm)	Depth		In-Plane Distance	
		MAE	Standard Deviation	MAE	Standard Deviation
Defective 1	9	0.34	0.29	0.41	0.11
	6	0.10	0.05	0.50	0.28
	3	0.16	0.05	0.26	0.16
Defective 2	9	0.33	0.15	0.28	0.15
	7	0.30	0.08	0.32	0.10
	6	0.23	0.17	0.44	0.14
	4	0.23	0.15	0.31	0.08
	3	0.36	0.06	0.40	0.16
Defective 3	6	-	-	0.27	0.18
Total		**0.26**	0.17	**0.37**	0.18

6.4.2.4 Visualisations

While the previous section quantitatively demonstrated the method's ability to localise defects in 3D, this capability is best appreciated through visualisations. Volumetric segmentation offers a significant advantage over image-based segmentation by providing comprehensive localisation information about defects. This enhanced localisation allows for more thorough reconstructions, which are particularly beneficial for constructing digital twins of components for testing or reporting purposes. To illustrate this benefit, Figure 71 presents 3D visual examples of the volumetric segmentation results, showcasing the model's performance across various defects and stepped samples. Despite the absence of TCG and significant variations in defect response levels at different thicknesses due to attenuation, the method manages

to deliver relatively consistent segmentation masks at different depths for defects of the same size by considering local temporal samples during inference. This resilience to variations in thickness and attenuation levels enhances the reliability and applicability of the segmentation method across a range of scenarios. After visualising the 3D segmented volumes, the opportunity of constructing digital twins becomes evident and highlights a clear benefit of the method.

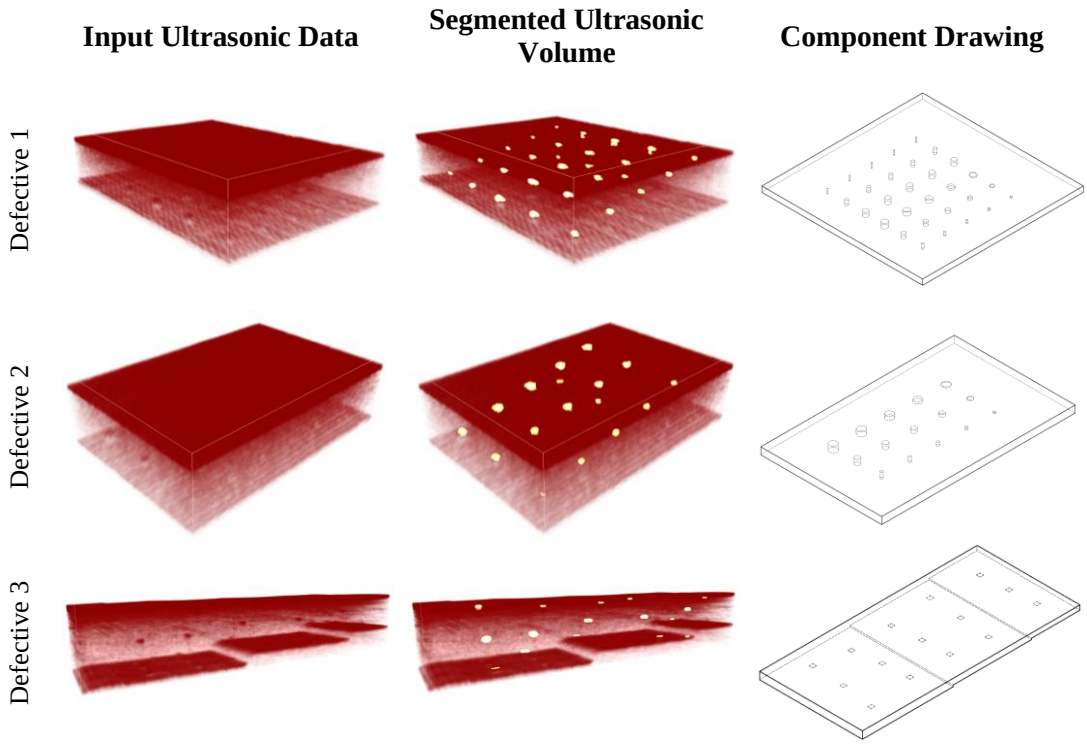


Figure 71: Visualisations of volumetric ultrasonic responses, their corresponding overlayed segmentations, and component drawing.

6.5 Industrial Demonstration

Given that this PhD research was sponsored by Spirit AeroSystems, it was important to investigate the applicability and efficacy of the proposed method when applied to industrial data and components. This sub-section gives an initial look into how the SSL method can be applied to data acquired from an industrial component and a different industrial acquisition setup.

An industrial sample was supplied by Spirit AeroSystems to test the performance of the SSL volumetric segmentation method. The sample is a stiffened panel, manufactured to BAPS 260-007, from the Learjet 85 with stringers which give rise to variable geometry. There are a range of PTFE insert which simulate inclusion/delamination defects. A 240 mm by 490 mm section of the component was scanned which contained 12 defects at a range of locations. A breakdown of the positions of the defects are given in Table 26.

Table 26: Summary of defects scanned in the Learjet 85 component.

Defect Position	Dimensions (mm) [Count]	Total
Skin (Near Tool/Top Surface)	20×10 [2], 10×10 [2], 5×5 [2]	3
Stringer (Near Tool/Top Surface)	20×10 [1], 10×10 [1], 5×5 [1]	3
Stringer (Near Bag/Bottom Surface)	20×10 [2], 10×10 [2], 5×5 [2]	6

Typical industrial data interpretation times for the Learjet sample can total 140 minutes; 80 minutes if the component is defect free, with a potential additional 60 minutes if the component contains defects. The component was scanned by Spirits' Belfast site using their Tecnatom phased array ultrasonic acquisition setup. A C-scan produced from the Tecnatom setup of the analysed section is given in Figure 73.

Data acquired at Spirits' Belfast site is usually analysed and stored post rectification, where the absolute value of the full waveform data is used to removed negative values. This rectification process compresses the data and removes phase information, similar to taking the envelope of the signal. However, unlike the enveloped signal the rectified signal gives an unsmooth response with multiple local maxima and minima for a given instantaneous response.

The SSL segmentation method is designed to be applied to smooth enveloped data and can highlight anomalous responses for given temporal signals. However, the method fails when applied to rectified data (as demonstrated in Figure 72). The presence of multiple maxima and minima for a given response makes it impractical to compare indexes over time, as minor temporal shifts can lead to significant changes in amplitude within a response. To address this issue the use of a moving average filter was explored. This can be applied to smooth the rectified signal, minimising the impact of multiple minima within a response. This approach has limitations such as requiring manually selecting the filter size to balance adequate smoothing and preserving the signal's response. However, it does provide a potential solution for the analysis of industrial data where the full waveform or enveloped signal is not available.

The results of applying the SSL method to the Learjet sample, based on data collected using the setup presented in Chapter 2, are presented herein (Figure 72). The evaluation focuses on the effects of different data processing techniques to ensure alignment with industrial data collection, storage, and interpretation. Specifically, the analysis includes the use of rectified data, moving-average rectified data, and Hilbert enveloped data. For this study, a minimum defect size filter was set to remove any defects below 3 mm in width, as consistent with the previous tests and a median threshold of 0.99999 was selected.

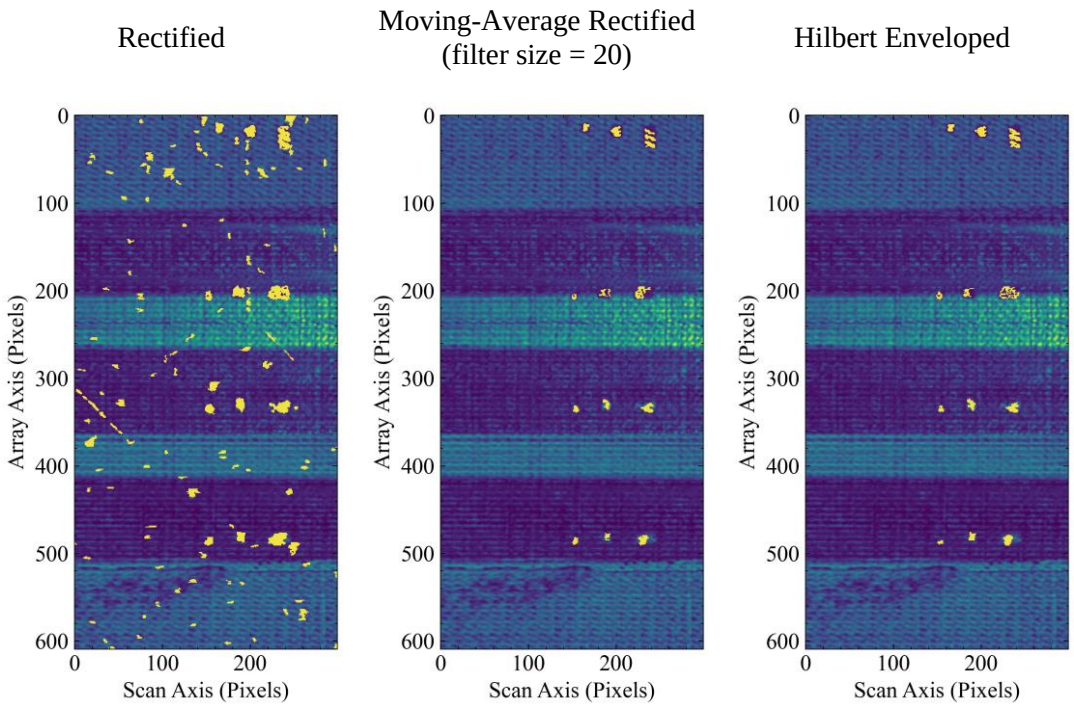


Figure 72: Impact on data processing for SSL method on Learjet data using the SEARCH lab acquisition system.

As qualitatively demonstrated though Figure 72, applying the SSL method directly to rectified data produces a significant number of false positives as expected. However, the moving average technique serves as an effective solution, enabling the method to be applied to industrial data where full waveform or enveloped data is not available.

To demonstrate transferability of the method to the acquisition system used by Spirit at their Belfast site, the method is applied to the moving average rectified data taken using the Tecnatom setup. The segmented C-scan for this is shown in Figure 73. Whilst there are additional false positives compared to the data acquired using the roller probe (on the Strathclyde SEARCH acquisition set up, Figure 72), each defect is clearly segmented. Despite the SSL model not having seen data from this acquisition setup, the results show promise for the method to be applied to both industrial parts and industrial acquisition systems. The results also demonstrate that in instances where the

Hilbert enveloped data is unavailable, applying a moving average to the rectified data can act as a practical workaround for the application of this methodology. It would be beneficial to conduct a wider study in future to assess the impact of different thresholds and moving average filter sizes. However, these results give promise to the application of this method to both industrial parts and acquisition setups.

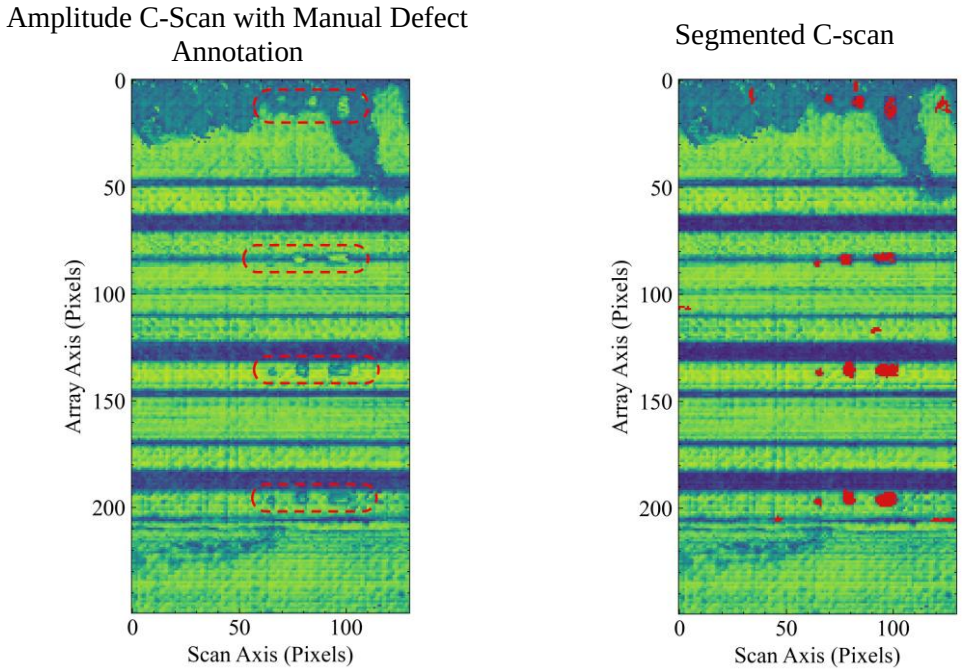


Figure 73: Amplitude C-scan of the Learjet component section acquired using the Tecnatom setup. Both manual defect annotation and SSL defect segmentations are shown.

Table 27: Comparison of correct defect detections and false indications for the different pre-processing and acquisition systems.

Acquisition System	Pre-processing	Correct Defect Detections	False Indications
SEARCH Lab Roller Probe	Rectified	12	101
	Moving Average	12	0
	Enveloped	12	0
Spirit Tecnatom	Moving Average	12	7

Table 27 highlights that although defects were consistently detected, different data pre-processing techniques significantly impacted the level of false indications. It also demonstrates that an appropriate moving average filter can match the performance of the enveloped data in this case. The method has shown promising transferability to Sprits' Tecnatom acquisition system, with the moving average filter clearly detecting all defects. Despite seven false positive indications, which could be due to a domain shift or due to the moving average filter not being specifically tuned for this acquisition system. This underscores a significant disadvantage of using this method with an additional pre-processing step: the introduction of a tuneable parameter due to the lack of full-waveform data. While a longer moving average filter kernel might help reduce false positives, it would also diminish defect response amplitudes (and potentially lead to missed defects). Therefore, it is preferable to apply the method to the enveloped data from the raw-waveform data as originally designed. In the future, acquiring full-waveform data using the Tecnatom acquisition system would be beneficial to determine if the additional false positives could be eliminated by using the enveloped data. It would also be beneficial to study the impact of changes in moving average filter size for when enveloped data is unavailable.

6.6 Conclusion

This chapter introduces a new approach for ultrasonic volumetric defect segmentation using SSL to address the need for defective training data. The method has been demonstrated for CFRP composites over different samples, defects, and geometries.

Volumetric segmentation provides information on defect sizing and localisation and allows for advanced visualisations which can facilitate the creation of digital twins for

reporting or testing. One of the biggest challenges when applying DL methods to NDE is the requirement for training data. This is compounded for volumetric segmentation tasks as volumetric training data is far less available and segmentation models typically require highly intensive labelling at a per voxel level.

This method employs a 1D CNN to learn the expected distributions for clean data along a given time index for a receptive element. This learned information is then used to generate a volumetric defect segmentation mask. Several advantages stem from this approach. Firstly, it circumvents the need for expensive and challenging-to-obtain labelled training data. By simplifying the pretext task to a 1D series, the accessibility of training data significantly increases; with millions of training samples derived from just four samples. Since this training data is defect-free, the limitation of obtaining large quantities of defective components no longer applies. Moreover, the size of the model used was only 1.94 MB. This means that although batching during inference is preferred for time-saving purposes, it is possible to reduce batch sizes or eliminate them altogether to allow the model to run on machines with more limited hardware.

Additionally, by framing the problem as one of anomaly prediction rather than positive defect identification, the method becomes more generalisable and robust to various types of defects, as long as they exhibit an amplitude response that deviates from the expected. This was demonstrated through testing on more challenging to detect PTFE inserts which were more accurately sized than the FBH. Furthermore, by treating the task as one of probabilistic sequence prediction, the segmentation approach becomes more interpretable compared to using a 3D end-to-end segmentation model. This aids

in demystifying some of the "black box" nature of DL predictions, which is a significant challenge in the NDE industry.

All defects were reliably detected across thresholds and processing steps. False positive indications were successfully eliminated for a 100% detection accuracy with the complete processing pipeline and a threshold of 0.9999999. The results showcased good in-plane and through-thickness localisation, exhibiting improvements over the previously presented fully supervised model [193] for in-plane localisation, with MAEs of 0.37 and 0.26 mm, respectively. Although through-thickness localisation performed slightly worse than previously demonstrated 3D U-Net model, it remains suitable for most applications, and it still falls below the in-plane localisation error. MAEs for defect sizing were also presented, revealing a negative correlation between sizing error and threshold. For a threshold of 0.9999999, the MAE aligned with the 6 dB drop at 1.41 mm. Due to the method's nature of detecting anomalous voxels in the ultrasonic domain, which may not always correspond directly to sizes in the physical domain, achieving accurate sizing remains challenging. However, the consistent over-sizing of defects enabled consistent calibration, which lead to a 57% reduction in sizing error.

The industrial case study demonstrated how the methodology could be successfully applied to an industrial part. It also highlighted a key challenge on how industrial data storage and representation can impact the applicability of the technique. However, the use of a moving average filter proved as a suitable solution to scenarios where the full waveform data may not be available to envelope using the Hilbert transform. Whilst further investigation would be beneficial regarding optimal moving average filter

lengths, along with testing on other components, the findings proved promising for the transferability of the approach to industrial components and data collection methods.

However, there are limitations to this method. Firstly, the method requires one consistent geometric axis to be applied to the data. While we demonstrated its robustness on stepped samples, more complex shapes may prove challenging, and the method would likely require modifications to accommodate such complexities. Secondly, most DL models shift the computational load of understanding a task to training, which can be done offline, prior to inference. This typically results in rapid inference results – a key benefit over other interpretation methods. However, the sequential nature of inference for this method means that there is an increased computational cost and time during interpretation. Although the full interpretation time was minimised to a maximum of 35 seconds using series batching for each B-scan, for less powerful machines, this may not always be feasible. Additionally, for larger parts, the inference time will increase as a result of longer scans. Despite this inference is still likely to be far quicker than human inspection in real-world applications.

In future work it would be interesting to evaluate this method against other, (perhaps simpler) methods of anomaly detection. It would also be beneficial to explore different models for the pre-text learning task and conduct a detailed analysis of how well the model can predict the distribution for the next consecutive value. Whilst the PNN gave results that performed well for segmentation, series forecasting is a significant, active area of research, and it is likely that a better approach exists to predict more accurate distributions of expected data points. A robust exploration of different approaches and how their accuracy impacts segmentation performance would be beneficial. It would

also be worthwhile exploring if this approach can be applied to different materials and scanning methodologies.

7 Summary and Future Work

7.1 Summary

The primary objective of this research was to develop DL solutions aimed at enhancing automation in the ultrasonic inspection of aerospace composites. To this end, the four chapters that form the core of the research presented in this thesis each contribute to a distinct aspect of applying DL to UT of composites. Supported closely by Spirit AeroSystems, this research holds significant potential for advancing automation within industrial settings.

Chapter 1 began by outlining the industrial motivation that contextualised this research. It underscored the growing significance of CFRP in the aerospace industry, which in turn highlighted the importance of NDE for ensuring component integrity. UT was identified as the most widely used NDE technique for inspecting aerospace composites. Advancements and the increasing uptake of robotics for sensor deployment has greatly progressed NDE automation. Despite this, the interpretation of UT results remained a largely manual process. This manual interpretation not only limits NDE automation, but also creates a bottleneck in the manufacturing process and introduces the potential for human error.

Chapter 2 provided an overview of background knowledge relevant to the thesis. It began with an introduction to the range of NDE techniques commonly used for composites, focusing on UT and its application in the aerospace industry. The use of phased arrays and mechanised scanning was also discussed, demonstrating how large-scale aerospace components can be inspected in industrial settings. The chapter also

explored the application of composites in aerospace and identified typical defects that can occur. Background information on AI, with an emphasis on CNNs was provided due to their importance in this research. Finally, the current use of ML in NDE was investigated, addressing key challenges such as data scarcity, problem definition, and model evaluation.

Chapter 3 addressed what is arguably the biggest challenge in applying ML to NDE: the scarcity of training data. To evaluate this, a CNN classifier was employed and four different methods for generating synthetic UT C-scans were investigated. Each method aimed to bridge the sim-to-real domain gap between simulated and real defect responses. Among these methods, one utilised a generative DL model, specifically CycleGAN. A modification to the model's loss function was introduced to constrain significant changes to defect responses while allowing the model to freely modify the noise response. Although the modified CycleGAN produced the most effective synthetic data, it was deduced that other methods performed nearly as well and might be preferable due to their robustness and ease of implementation. Guided Grad-CAM was employed to compare models trained on synthetic and experimental data. While this provided only a qualitative analysis, it suggested that models trained on synthetic data use similar features for classification as those trained on experimental data. This finding reinforces the potential of synthetic data to mitigate the issue of training data scarcity and helps to limit the additional interpretability concern which may occur from the use of synthetic training data. This work proved fundamental to the subsequent chapters 4 and 5, as it established a robust framework for generating synthetic training data, which was crucial for the development and training of more advanced machine learning models.

Chapter 4 delved into the potential of harnessing the complete 3D ultrasonic volume for defect detection. While human operators typically visualise this volume as images for easier interpretation, ML systems are not confined to such 2D representations. The underlying intuition was that by refraining from compressing the time or spatial domain during imaging, the ML model could glean richer information to assess defect responses. This endeavour necessitated the generation of synthetic volumetric datasets for training, building upon the A-scan noise generation method introduced in previous chapters. This method, initially designed for generating synthetic C-scan images, was extended to cater to volumetric datasets. A NAS was conducted, leading to the discovery of an optimal architecture that was then evaluated against two other models. To further bridge the gap between synthetic and experimental domains, two domain-specific augmentation methods were introduced. These methods significantly enhanced the classification performance. The study conclusively demonstrated the feasibility of training a DL model to effectively detect defects using the complete volumetric ultrasonic data.

Chapter 5 advanced beyond the scope of the previous chapter by focusing on defect segmentation within the ultrasonic volume. This expanded approach provided valuable additional information for defect characterisation, such as defect sizing along with in-plane and through-thickness localisation. These aspects are crucial for tasks such as conducting repairs or constructing digital twins, along with determining the criticality of defects. To achieve this, a modified 3D U-Net was trained in a fully supervised manner using synthetic defect responses and their corresponding defect masks. Evaluation of the localisation performance was conducted against known defect depth and 6 dB centroid, yielding a MAE of 0.08 mm and 0.57 mm, respectively. Sizing

accuracy was compared to the 6 dB drop criterion. Initially, there was a 55% increase in error compared to the 6 dB drop criterion. However, after applying a correction factor to account for the disparity between the experimental and synthetic domains, a 35% reduction in error was achieved over the 6 dB drop method in defect sizing. Furthermore, the model was subjected to testing on out-of-distribution defect responses from PTFE inserts. This testing revealed a 71% increase in defect sizing error compared to in-distribution defects. This underscores a notable limitation of supervised training, which necessitates adequate coverage of the test distribution of defects during training. Despite the utility of synthetic data in mitigating data scarcity, modelling the full range of defects and the inherent variability within defect types, as necessary for wide generalisability, remains a challenge in NDE.

Chapter 6 furthered the research into segmentation of the full ultrasonic volume by developing a self-supervised approach for training a segmentation model. In contrast to chapter 5, which employed supervised learning, this method did not require examples of defect responses during training and only used defect free data. This represents a significant advantage as experimental defect-free data is much more readily available. As noted in the previous chapter, fully labelled experimental responses are rarely available, and accurately simulating the range of defect responses required for training a supervised model is a substantial challenge. The SSL model demonstrated good results for in-plane and through-thickness localisation, with MAEs of 0.26 mm and 0.37 mm, respectively. Since the SSL method is not trained to identify true defect size in the physical domain from the ultrasonic domain, the initial MAE for defect sizing was greater than that achieved with fully supervised training. However, due to the consistency of the method, it is possible to calibrate to physical defect sizes.

This calibration resulted in an MAE of less than the element pitch, specifically 0.58 mm. Furthermore, by treating the segmentation problem as one of anomaly detection the method generalises well to different defect responses as demonstrated with testing on PTFE inserts as well as FBHs. The effectiveness of the method was demonstrated on an industrial component. Where industrial data collection limits the availability of enveloped data processing, it was evidenced that an appropriate moving average filter could be applied to allows the applicability of this method. This approach simultaneously solves the problem of needing defect responses during training, allows for full 3D segmentation, and demonstrates impressive generalisability.

7.2 Future Work and Final Remarks

There remain challenges in applying DL to NDE. Many valuable lessons have been learnt during this research and are useful for future research or industrial DL NDE tasks. The author believes that, in general, detection problems should be framed as issues of anomaly detection and addressed using unsupervised or self-supervised methods. This is primarily due to the lack of large training datasets available for supervised training and improved generalisability. It has been demonstrated that even defect segmentation, which traditionally relies on supervised learning, can be reframed in this manner with significant benefits. In the future, large open-source datasets may become available, facilitating the development of generalisable supervised models. However, until this happens, alternative forms of training are likely to be more robust to the variability encountered in NDE and even if these were to become available may still be inferior to supervised training methods.

There is still scope for further investigation into the research presented in this thesis. The most noteworthy future work would be the incorporation of a wide range of

naturally occurring defects into the test sets to evaluate model generalisability. As observed in Chapter 5, model performance can decrease when exposed to samples outside of the training distribution. Before deploying DL-based solutions in industry, it is crucial to robustly test them on a variety of naturally occurring defects. If such testing were conducted, one would likely see a decrease in performance for the presented models trained in a supervised manner, due to the lack of extensive training data.

Chapters 3-5 demonstrated that synthetic data can be effective for training models. However, to produce generalisable models for industrial applications, it would likely require generating a much larger synthetic training set that encompasses a broader range of variability seen in naturally occurring defects. This would necessitate the accurate simulation of a greater variety of defect responses, encompassing different classes of defects as well as greater intra-class variability. In addition to variability in defects, the base sample should also be diversified, to account for differences in components such as ply layup and geometric changes. Should these datasets become available, much of the work conducted in Chapters 4-5 could be revisited, with retraining and evaluation to cover a wider spectrum of defects and component variability.

Chapters 5 and 6 demonstrated how volumetric segmentation could be used for defect characterisation. However, for complete characterisation, a key piece of information is missing: the class of defect. Multi-class classification was not explored in this work primarily due to the lack of available test data. The volumetric defect detection model presented in Chapter 4 could be easily modified and retrained to provide multi-class

predictions. Here, the additional information available by analysing the full volume may prove beneficial in improving classification accuracy over 2D image analysis. To train such a model, an expanded synthetic (or experimental) training dataset with associated class labels would be required. Alternatively, SSL approaches such as Siamese networks [198] could be effective if large labelled training datasets are unavailable. However, any approach would still necessitate a test set with multiple defect classes for evaluation.

Whilst chapter 6 demonstrated the potential of the methodology on industrial components, further exploration would be beneficial. Industrial scans are likely to vary in quality and sensing methodologies, which may necessitate further research. Real components can vary significantly in geometry, making automated detection more challenging, as geometrical responses can often be indistinguishable from defect responses without additional information. While Chapter 6 demonstrated the effectiveness of the SSL model on stepped samples and an industrial sample with stringers, the model necessitated inference parallel to any change in thickness. Industrial components with more complex geometries could limit its applicability or require modification of the methodology.

In addition to testing on a greater variety of industrial components, there is also a regulatory factor to consider for industrial adoption. Further work would be beneficial to assess how these tools and systems could be used in industry and what is required to approve them for use. Regulatory approval may become the biggest barrier to the uptake and use of these techniques, necessitating comprehensive evaluations to ensure compliance with industry standards and regulations. Addressing these challenges will

be crucial for the successful deployment of these advanced methodologies in real-world industrial settings.

Initial work has been undertaken to build on this DL research to produce automated UT software solutions with robotic integration. Figure 74 shows a screenshot of the developed software and its integration with a collaborative robot-based sensor delivery platform. This system was recently presented at the BINDT Aerospace Conference [199]. However, there remains significant potential for further development. This could involve collaborating with NDE operators to create more specialised and effective human-in-the-loop automation tools, designed with operator feedback. The integration with robotic systems paves the way for future advancements in automation, including the ability to automatically re-scan or perform rework on components. By taking a holistic approach, incorporating operator insights, and advancing robotic integration, the system could become more adaptive and efficient, leading to improved defect detection and repair processes in industrial applications.

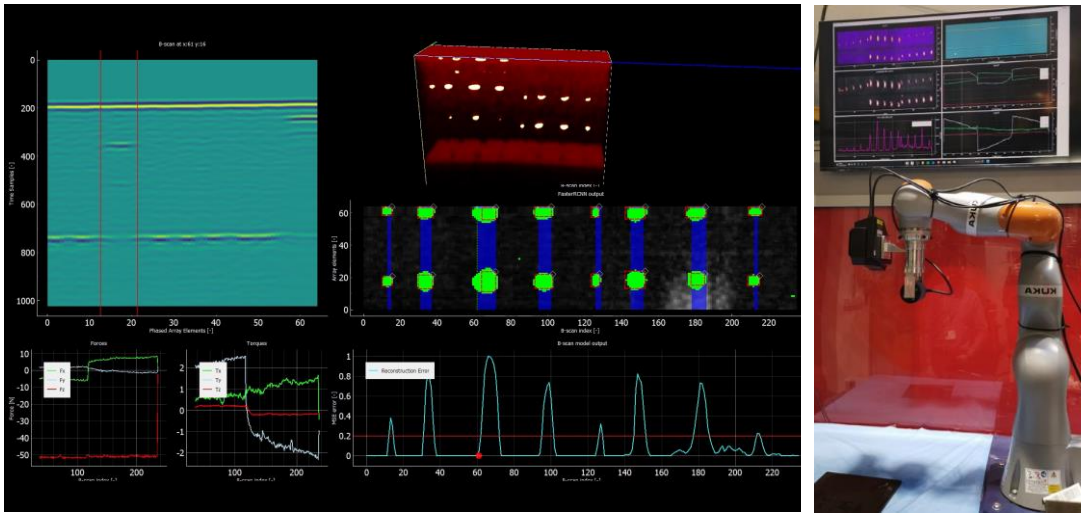


Figure 74: Screenshot of the graphical user interface developed to interact with DL model outputs from UT scans (left) and demonstration of the system integrated into a flexible robotic scanning system (right).

8 Bibliography

[1] J. Zhang, G. Lin, U. Vaidya, and H. Wang, 'Past, present and future prospective of global carbon fibre composite developments and applications', *Composites Part B: Engineering*, vol. 250, p. 110463, Feb. 2023, doi: 10.1016/j.compositesb.2022.110463.

[2] J. Skoczylas, S. Samborski, and M. Kłonica, 'THE APPLICATION OF COMPOSITE MATERIALS IN THE AEROSPACE INDUSTRY', *Journal of Technology and Exploitation in Mechanical Engineering*, vol. 5, Jan. 2019, doi: 10.35784/jteme.73.

[3] B. Parveez, M. I. Kittur, I. A. Badruddin, S. Kamangar, M. Hussien, and M. A. Umarfarooq, 'Scientific Advancements in Composite Materials for Aircraft Applications: A Review', *Polymers (Basel)*, vol. 14, no. 22, p. 5007, Nov. 2022, doi: 10.3390/polym14225007.

[4] J. Jodhani, A. Handa, A. Gautam, Ashwni, and R. Rana, 'Ultrasonic non-destructive evaluation of composites: A review', *Materials Today: Proceedings*, vol. 78, pp. 627–632, Jan. 2023, doi: 10.1016/j.matpr.2022.12.055.

[5] R. Henrich and U. Schnars, 'Applications of NDT Methods on Composite Structures in Aerospace Industry', *e-Journal of Nondestructive Testing*, vol. 11, no. 12, Dec. 2006, Accessed: Jun. 19, 2024. [Online]. Available: https://www.ndt.net/search/docs.php3?id=4180&msgID=0&rootID=0

[6] S. Cumblidge, A. D'Agostino, S. Morrow, C. Franklin, and N. Nughes, 'Review of Human Factors Research in Nondestructive Examination.', in *US Nuclear Regulatory Commission-Pacific Northwest National Laboratory 7th European-American Workshop on Reliability of NDE.*, Feb. 2017.

[7] P. Gardner *et al.*, 'Machine learning at the interface of structural health monitoring and non-destructive evaluation', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 378, no. 2182, p. 20190581, Oct. 2020, doi: 10.1098/rsta.2019.0581.

[8] C. Mineo *et al.*, 'Robotic Geometric and Volumetric Inspection of High Value and Large Scale Aircraft Wings', in *2019 IEEE 5th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, Jun. 2019, pp. 82–86. doi:

10.1109/MetroAeroSpace.2019.8869667.

[9] L. Séguin-Charbonneau, J. Walter, L.-D. Théroux, L. Scheed, A. Beausoleil, and B. Masson, ‘Automated defect detection for ultrasonic inspection of CFRP aircraft components’, *NDT & E International*, vol. 122, p. 102478, Sep. 2021, doi: 10.1016/j.ndteint.2021.102478.

[10] ‘Introduction to non-destructive testing’, Aerospace Testing International. Accessed: Nov. 17, 2021. [Online]. Available: https://www.aerospacetestinginternational.com/features/introduction-to-non-destructive-testing.html

[11] F. W. Margrave, K. Rigas, D. A. Bradley, and P. Barrowcliffe, ‘The use of neural networks in ultrasonic flaw detection’, *Measurement*, vol. 25, no. 2, pp. 143–154, Mar. 1999, doi: 10.1016/S0263-2241(98)00075-X.

[12] J. Ye, S. Ito, and N. Toyama, ‘Computerized Ultrasonic Imaging Inspection: From Shallow to Deep Learning’, *Sensors (Basel)*, vol. 18, no. 11, p. 3820, Nov. 2018, doi: 10.3390/s18113820.

[13] B. Valeske, A. Osman, F. Römer, and R. Tschuncky, ‘Next Generation NDE Sensor Systems as IIoT Elements of Industry 4.0’, *Research in Nondestructive Evaluation*, vol. 31, no. 5–6, pp. 340–369, Nov. 2020, doi: 10.1080/09349847.2020.1841862.

[14] S. Cantero-Chinchilla, P. D. Wilcox, and A. J. Croxford, ‘Deep learning in automated ultrasonic NDE -- developments, axioms and opportunities’, *arXiv:2112.06650 [eess]*, Dec. 2021, Accessed: Jan. 12, 2022. [Online]. Available: http://arxiv.org/abs/2112.06650

[15] S. Gholizadeh, ‘A review of non-destructive testing methods of composite materials’, *Procedia Structural Integrity*, vol. 1, pp. 50–57, 2016, doi: 10.1016/j.prostr.2016.02.008.

[16] M. V. Felice and Z. Fan, ‘Sizing of flaws using ultrasonic bulk wave testing: A review’, *Ultrasonics*, vol. 88, pp. 26–42, Aug. 2018, doi: 10.1016/j.ultras.2018.03.003.

[17] S. C. Olisa, M. A. Khan, and A. Starr, ‘Review of Current Guided Wave Ultrasonic Testing (GWUT) Limitations and Future Directions’, *Sensors (Basel)*, vol. 21, no. 3, p. 811, Jan. 2021, doi: 10.3390/s21030811.

[18] C. Meola, S. Boccardi, G. M. Carlomagno, N. D. Boffa, E. Monaco, and F. Ricci, 'Nondestructive evaluation of carbon fibre reinforced composites with infrared thermography and ultrasonics', *Composite Structures*, vol. 134, pp. 845–853, Dec. 2015, doi: 10.1016/j.compstruct.2015.08.119.

[19] D. K. Hsu, '15 - Non-destructive evaluation (NDE) of aerospace composites: ultrasonic techniques', in *Non-Destructive Evaluation (NDE) of Polymer Matrix Composites*, V. M. Karbhari, Ed., in Woodhead Publishing Series in Composites Science and Engineering. , Woodhead Publishing, 2013, pp. 397–422. doi: 10.1533/9780857093554.3.397.

[20] 'Ultrasonic Test - an overview | ScienceDirect Topics'. Accessed: Jun. 14, 2024. [Online]. Available: https://www.sciencedirect.com/topics/physics-and-astronomy/ultrasonic-test

[21] I. Papa, V. Lopresto, and A. Langella, 'Ultrasonic inspection of composites materials: Application to detect impact damage', *International Journal of Lightweight Materials and Manufacture*, vol. 4, no. 1, pp. 37–42, Mar. 2021, doi: 10.1016/j.ijlmm.2020.04.002.

[22] C. Holmes, B. W. Drinkwater, and P. D. Wilcox, 'Post-processing of the full matrix of ultrasonic transmit–receive array data for non-destructive evaluation', *NDT & E International*, vol. 38, no. 8, pp. 701–711, Dec. 2005, doi: 10.1016/j.ndteint.2005.04.002.

[23] S. Moon *et al.*, 'FEA-Based Ultrasonic Focusing Method in Anisotropic Media for Phased Array Systems', *Applied Sciences*, vol. 11, no. 19, Art. no. 19, Jan. 2021, doi: 10.3390/app11198888.

[24] C. Aguado, 'Ultrasonic Techniques and Industrial Robots: Natural Evolution of Inspection Systems'.

[25] D. Garnier, D. Garnier, P. Louviot, and A. Tachattahte, 'Robotised UT Transmission NDT of Composite Complex Shaped Parts'.

[26] C. Mineo *et al.*, 'Flexible integration of robotics, ultrasonics and metrology for the inspection of aerospace components', *AIP Conference Proceedings*, vol. 1806, no. 1, p. 020026, Feb. 2017, doi: 10.1063/1.4974567.

[27] R. A. Smith and B. Clarke, 'Ultrasonic C-scan determination of ply stacking sequence in carbon-fibre composites.', *Insight - Non-Destructive Testing and*

Condition Monitoring, vol. 36, no. 10, pp. 741–747, 1994.

[28] P. Vaara and J. Leinonen, 'Technology Survey on NDT of Carbon-fiber Composites', 2012. Accessed: Feb. 26, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Technology-Survey-on-NDT-of-Carbon-fiber-Composites-Vaara-Leinonen/7d8b35f344c13149b177995ed210431f5d36dff0

[29] B. Djordjevic, 'Non Destructive Test Technology for the Composite', p. 7, Jan. 2009.

[30] D. Moore and A. M.-E. Dorado, 'Composite Material Characterization using Acoustic Wave Speed Measurements', presented at the ASNT Annual Conference, 2015, p. 5. [Online]. Available: https://www.osti.gov/biblio/1326353

[31] Ley, O. and V. Godinez, 'Non-destructive evaluation (NDE) of aerospace composites: application of infrared (IR) thermography', doi: 10.1533/9780857093554.3.309.

[32] A. Kokurov and D. Subbotin, 'Ultrasonic detection of manufacturing defects in multilayer composite structures', *IOP Conference Series: Materials Science and Engineering*, vol. 1023, p. 012013, Jan. 2021, doi: 10.1088/1757-899X/1023/1/012013.

[33] F. Heinecke and C. Willberg, 'Manufacturing-Induced Imperfections in Composite Parts Manufactured via Automated Fiber Placement', *J. Compos. Sci.*, vol. 3, no. 2, p. 56, Jun. 2019, doi: 10.3390/jcs3020056.

[34] M. Jolly *et al.*, 'Review of Non-destructive Testing (NDT) Techniques and their Applicability to Thick Walled Composites', *Procedia CIRP*, vol. 38, pp. 129–136, Jan. 2015, doi: 10.1016/j.procir.2015.07.043.

[35] H. Taheri and A. A. Hassen, 'Nondestructive Ultrasonic Inspection of Composite Materials: A Comparative Advantage of Phased Array Ultrasonic', *Applied Sciences*, vol. 9, no. 8, Art. no. 8, Jan. 2019, doi: 10.3390/app9081628.

[36] S. Bayat, A. Jamzad, N. Zobeiry, A. Poursartip, P. Mousavi, and P. Abolmaesumi, 'Temporal enhanced Ultrasound: A new method for detection of porosity defects in composites', *Composites Part A: Applied Science and Manufacturing*, vol. 164, p. 107259, Jan. 2023, doi: 10.1016/j.compositesa.2022.107259.

[37] G. Fernlund, J. Wells, L. Fahrang, J. Kay, and A. Poursartip, 'Causes and

remedies for porosity in composite manufacturing', *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 139, p. 012002, Jul. 2016, doi: 10.1088/1757-899X/139/1/012002.

[38] K. Senthil, A. Arockiarajan, R. Palaninathan, B. Santhosh, and K. M. Usha, 'Defects in composite structures: Its effects and prediction methods – A comprehensive review', *Composite Structures*, vol. 106, pp. 139–149, Dec. 2013, doi: 10.1016/j.compstruct.2013.06.008.

[39] R. Telford, A. O'Carroll, R. S. Pierce, and T. M. Young, 'A novel method to produce kiss-bonds in composites components for NDI and characterisation purposes', *Composites Part B: Engineering*, vol. 173, p. 106926, Sep. 2019, doi: 10.1016/j.compositesb.2019.106926.

[40] A. Poudel, S. S. Shrestha, J. S. Sandhu, T. P. Chu, and C. G. Pergantis, 'Comparison and analysis of Acoustography with other NDE techniques for foreign object inclusion detection in graphite epoxy composites', *Composites Part B: Engineering*, vol. 78, pp. 86–94, Sep. 2015, doi: 10.1016/j.compositesb.2015.03.048.

[41] H. M. Hsiao and I. M. Daniel, 'Effect of fiber waviness on stiffness and strength reduction of unidirectional composites under compressive loading', *Composites Science and Technology*, vol. 56, no. 5, pp. 581–593, Jan. 1996, doi: 10.1016/0266-3538(96)00045-0.

[42] G. Karami and M. Garnich, 'Effective moduli and failure considerations for composites with periodic fiber waviness', *Composite Structures*, vol. 67, no. 4, pp. 461–475, Mar. 2005, doi: 10.1016/j.compstruct.2004.02.005.

[43] C. Shah, S. Bosse, and A. von Hehl, 'Taxonomy of Damage Patterns in Composite Materials, Measuring Signals, and Methods for Automated Damage Diagnostics', *Materials (Basel)*, vol. 15, no. 13, p. 4645, Jul. 2022, doi: 10.3390/ma15134645.

[44] K. Ajay, 'Best Practice Guide Non-Destructive Testing of Composites'. National Composites Network, 2013. [Online]. Available: https://avaloncsl.files.wordpress.com/2013/01/ncn-best-practice-ndt.pdf

[45] R. H. Bossi and G. E. Georgeson, '16 - Nondestructive testing of aerospace composites', in *Polymer Composites in the Aerospace Industry (Second Edition)*, P. Irving and C. Soutis, Eds., in Woodhead Publishing Series in Composites

Science and Engineering. , Woodhead Publishing, 2020, pp. 461–489. doi: 10.1016/B978-0-08-102679-3.00016-2.

[46] W. Ertel, *Introduction to Artificial Intelligence*. in Undergraduate Topics in Computer Science. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-58487-4.

[47] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, and A. Rahmim, 'A Brief History of AI: How to Prevent Another Winter (A Critical Review)', *PET Clin*, vol. 16, no. 4, pp. 449–469, Oct. 2021, doi: 10.1016/j.cpet.2021.07.001.

[48] G. F. Luger and W. A. Stubblefield, *Artificial Intelligence and the Design of Expert Systems*. Benjamin/Cummings Publishing Company, 1989.

[49] M. Campbell, A. J. Hoane, and F. Hsu, 'Deep Blue', *Artificial Intelligence*, vol. 134, no. 1, pp. 57–83, Jan. 2002, doi: 10.1016/S0004-3702(01)00129-1.

[50] J. Weizenbaum, 'ELIZA—a computer program for the study of natural language communication between man and machine', *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, doi: 10.1145/365153.365168.

[51] J. M. Mira, 'Symbols versus connections: 50 years of artificial intelligence', *Neurocomputing*, vol. 71, no. 4, pp. 671–680, Jan. 2008, doi: 10.1016/j.neucom.2007.06.009.

[52] Y. Shang, '5 - Expert Systems', in *The Electrical Engineering Handbook*, W.-K. Chen, Ed., Burlington: Academic Press, 2005, pp. 367–377. doi: 10.1016/B978-012170960-0/50031-1.

[53] P. Smolensky, 'Connectionist AI, symbolic AI, and the brain', *Artif Intell Rev*, vol. 1, no. 2, pp. 95–109, Jun. 1987, doi: 10.1007/BF00130011.

[54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, 'Learning representations by back-propagating errors', *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[55] I. H. Sarker, 'AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems', *Sn Computer Science*, vol. 3, no. 2, p. 158, Feb. 2022, doi: 10.1007/s42979-022-01043-x.

[56] M. Elahi, S. O. Afolaranmi, J. L. Martinez Lastra, and J. A. Perez Garcia, 'A comprehensive literature review of the applications of AI techniques through the lifecycle of industrial equipment', *Discov Artif Intell*, vol. 3, no. 1, p. 43, Dec.

2023, doi: 10.1007/s44163-023-00089-x.

[57] K. Ofosu-Ampong, 'Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions', *Telematics and Informatics Reports*, vol. 14, p. 100127, Jun. 2024, doi: 10.1016/j.teler.2024.100127.

[58] S. A. Alowais *et al.*, 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice', *BMC Medical Education*, vol. 23, no. 1, p. 689, Sep. 2023, doi: 10.1186/s12909-023-04698-z.

[59] D. Khurana, A. Koli, K. Khatter, and S. Singh, 'Natural language processing: state of the art, current trends and challenges', *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[60] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.

[61] W. S. McCulloch and W. Pitts, 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.

[62] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[63] F. Rosenblatt, 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.

[64] K. Hornik, M. Stinchcombe, and H. White, 'Multilayer feedforward networks are universal approximators', *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[65] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[66] Y. Bengio, P. Simard, and P. Frasconi, 'Learning long-term dependencies with gradient descent is difficult', *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[67] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[68] L. Ciampiconi, A. Elwood, M. Leonardi, A. Mohamed, and A. Rozza, ‘A survey and taxonomy of loss functions in machine learning’, Jan. 13, 2023, *arXiv*: arXiv:2301.05579. doi: 10.48550/arXiv.2301.05579.

[69] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, ‘A Comprehensive Survey of Loss Functions in Machine Learning’, *Ann. Data. Sci.*, vol. 9, no. 2, pp. 187–212, Apr. 2022, doi: 10.1007/s40745-020-00253-5.

[70] D. P. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimization’, Jan. 29, 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

[71] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, ‘Densely Connected Convolutional Networks’, *arXiv:1608.06993 [cs]*, Jan. 2018, Accessed: Oct. 04, 2021. [Online]. Available: http://arxiv.org/abs/1608.06993

[72] R. Abdulkadirov, P. Lyakhov, and N. Nagornov, ‘Survey of Optimization Algorithms in Modern Neural Networks’, *Mathematics*, vol. 11, no. 11, Art. no. 11, Jan. 2023, doi: 10.3390/math11112466.

[73] R. Sun, ‘Optimization for deep learning: theory and algorithms’, Dec. 18, 2019, *arXiv*: arXiv:1912.08957. doi: 10.48550/arXiv.1912.08957.

[74] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, ‘Deep learning in computer vision: A critical review of emerging techniques and application scenarios’, *Machine Learning with Applications*, vol. 6, p. 100134, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.

[75] N. O’Mahony *et al.*, ‘Deep Learning vs. Traditional Computer Vision’, in *Advances in Computer Vision*, vol. 943, K. Arai and S. Kapoor, Eds., in Advances in Intelligent Systems and Computing, vol. 943. , Cham: Springer International Publishing, 2020, pp. 128–144. doi: 10.1007/978-3-030-17795-9_10.

[76] L. Liu *et al.*, ‘Deep Learning for Generic Object Detection: A Survey’, *Int J Comput Vis*, vol. 128, no. 2, pp. 261–318, Feb. 2020, doi: 10.1007/s11263-019-01247-4.

[77] S. Albelwi and A. Mahmood, ‘A Framework for Designing the Architectures of Deep Convolutional Neural Networks’, *Entropy*, vol. 19, no. 6, Art. no. 6, Jun. 2017, doi: 10.3390/e19060242.

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘ImageNet Classification with Deep Convolutional Neural Networks’, in *Advances in Neural Information*

Processing Systems, Curran Associates, Inc., 2012. Accessed: Feb. 27, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[79] 'ImageNet'. Accessed: Jan. 29, 2022. [Online]. Available: https://image-net.org/

[80] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[81] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[82] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', Mar. 02, 2022, *arXiv*: arXiv:2201.03545. Accessed: Jun. 16, 2023. [Online]. Available: http://arxiv.org/abs/2201.03545

[83] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

[84] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, 'A survey on deep learning techniques for image and video semantic segmentation', *Applied Soft Computing*, vol. 70, pp. 41–65, Sep. 2018, doi: 10.1016/j.asoc.2018.05.018.

[85] A. Ghosh, N. D. Jana, S. Das, and R. Mallipeddi, 'Two-Phase Evolutionary Convolutional Neural Network Architecture Search for Medical Image Classification', *IEEE Access*, vol. 11, pp. 115280–115305, 2023, doi: 10.1109/ACCESS.2023.3323705.

[86] 'CIFAR-10 and CIFAR-100 datasets'. Accessed: Jan. 29, 2022. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[87] X. Zhu, C. Vondrick, C. Fowlkes, and D. Ramanan, 'Do We Need More Training Data?', *Int J Comput Vis*, vol. 119, no. 1, pp. 76–92, Aug. 2016, doi: 10.1007/s11263-015-0812-2.

[88] C. Shorten and T. M. Khoshgoftaar, 'A survey on Image Data Augmentation for Deep Learning', *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi:

10.1186/s40537-019-0197-0.

[89] Y. Sun, P. Bai, H. Sun, and P. Zhou, ‘Real-time automatic detection of weld defects in steel pipe’, *NDT & E International*, vol. 38, no. 7, pp. 522–528, Oct. 2005, doi: 10.1016/j.ndteint.2005.01.011.

[90] A. A. Markov, D. A. Shpagin, and M. N. Shilov, ‘Ultrasonic Multichannel Flaw Detector for Testing Rails with Signal Recording’, *Russian Journal of Nondestructive Testing*, vol. 39, no. 2, pp. 105–114, Feb. 2003, doi: 10.1023/B:RUNT.0000008386.07674.97.

[91] A. Osman, Y. Duan, and V. Kaftandjian, ‘Applied Artificial Intelligence in NDE’, in *Handbook of Nondestructive Evaluation 4.0*, N. Meyendorf, N. Ida, R. Singh, and J. Vrana, Eds., Cham: Springer International Publishing, 2021, pp. 1–35. doi: 10.1007/978-3-030-48200-8_49-1.

[92] S. Barut, V. Bissauge, G. Ithurralde, and W. Claassens, ‘Computer-aided analysis of ultrasound data to speed-up the release of aerospace CFRP components’, *e-Journal of Nondestructive Testing*, vol. 17, no. 07, Jul. 2012, [Online]. Available: https://www.ndt.net/search/docs.php3?id=12429&msgID=0&rootID=0

[93] J. Aldrin, C. Coughlin, D. Forsyth, and J. Welter, ‘Progress on the Development of Automated Data Analysis Algorithms and Software for Ultrasonic Inspection of Composites’, *AIP Conference Proceedings*, vol. 1581, Jan. 2014, doi: 10.1063/1.4865058.

[94] D. Guo, G. Jiang, X. Lin, and Y. Wu, ‘Automated ultrasonic testing for 3D laser-rapid prototyping blisk blades’, in *2016 7th International Conference on Mechanical and Aerospace Engineering (ICMAE)*, Jul. 2016, pp. 214–218. doi: 10.1109/ICMAE.2016.7549537.

[95] S. J. Song and L. W. Schmerr, ‘Ultrasonic flaw classification in weldments using probabilistic neural networks’, *J Nondestruct Eval*, vol. 11, no. 2, pp. 69–77, Jun. 1992, doi: 10.1007/BF00568290.

[96] S.-J. Song, H.-J. Kim, and H. Cho, ‘Development of an intelligent system for ultrasonic flaw classification in weldments’, *Nuclear Engineering and Design*, vol. 212, no. 1, pp. 307–320, Mar. 2002, doi: 10.1016/S0029-5493(01)00495-2.

[97] N. Munir, H.-J. Kim, J. Park, S.-J. Song, and S.-S. Kang, ‘Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions’,

Ultrasonics, vol. 94, pp. 74–81, Apr. 2019, doi: 10.1016/j.ultras.2018.12.001.

[98] P. Zacharis, G. West, G. Dobie, T. Lardner, and A. Gachagan, 'Data-Driven Analysis of Ultrasonic Inspection Data of Pressure Tubes', *Nuclear Technology*, vol. 202, no. 2–3, pp. 153–160, Jun. 2018, doi: 10.1080/00295450.2017.1421803.

[99] R. J. Pyle, R. L. T. Bevan, R. R. Hughes, A. A. S. Ali, and P. D. Wilcox, 'Domain Adapted Deep-Learning for Improved Ultrasonic Crack Characterization Using Limited Experimental Data', *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1485–1496, Apr. 2022, doi: 10.1109/TUFFC.2022.3151397.

[100] I. Virkkunen, T. Koskinen, O. Jessen-Juhler, and J. Rinta-aho, 'Augmented Ultrasonic Data for Machine Learning', *J Nondestruct Eval*, vol. 40, no. 1, p. 4, Jan. 2021, doi: 10.1007/s10921-020-00739-5.

[101] D. Medak, L. Posilovic, M. Subasic, M. Budimir, and S. Loncaric, 'Automated Defect Detection From Ultrasonic Images Using Deep Learning', *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, vol. 68, no. 10, pp. 3126–3134, Oct. 2021, doi: 10.1109/TUFFC.2021.3081750.

[102] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation', *Nat Methods*, vol. 18, no. 2, Art. no. 2, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[103] D. Medak, L. Posilović, M. Subašić, M. Budimir, and S. Lončarić, 'DefectDet: A deep learning architecture for detection of defects with extreme aspect ratios in ultrasonic images', *Neurocomputing*, vol. 473, pp. 107–115, Feb. 2022, doi: 10.1016/j.neucom.2021.12.008.

[104] J. Ye and N. Toyama, 'Benchmarking Deep Learning Models for Automatic Ultrasonic Imaging Inspection', *IEEE Access*, vol. 9, pp. 36986–36994, 2021, doi: 10.1109/ACCESS.2021.3062860.

[105] A. Rytter, 'Vibrational Based Inspection of Civil Engineering Structures', 1993.

[106] R. J. Pyle, R. L. T. Bevan, R. R. Hughes, R. K. Rachev, A. A. S. Ali, and P. D. Wilcox, 'Deep Learning for Ultrasonic Crack Characterization in NDE', *IEEE*

Trans. Ultrason., Ferroelect., Freq. Contr., vol. 68, no. 5, pp. 1854–1865, May 2021, doi: 10.1109/TUFFC.2020.3045847.

[107] A. Wirgin, 'The inverse crime', *arXiv:math-ph/0401050*, Jan. 2004, Accessed: Jan. 06, 2022. [Online]. Available: http://arxiv.org/abs/math-ph/0401050

[108] S. Ali *et al.*, 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', *Information Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[109] S. A. and S. R., 'A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends', *Decision Analytics Journal*, vol. 7, p. 100230, Jun. 2023, doi: 10.1016/j.dajour.2023.100230.

[110] M. Abdar *et al.*, 'A review of uncertainty quantification in deep learning: Techniques, applications and challenges', *Information Fusion*, vol. 76, pp. 243–297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.

[111]A. Young *et al.*, 'Capturing Symbolic Expert Knowledge for the Development of Industrial Fault Detection Systems: Manual and Automated Approaches', *International journal of COMADEM*.

[112] 'EASA Artificial Intelligence Roadmap 1.0', EASA. Accessed: Jan. 12, 2022. [Online]. Available: https://www.easa.europa.eu/document-library/general-publications/easa-artificial-intelligence-roadmap-10

[113] 'EASA releases its Concept Paper "First usable guidance for Level 1 machine learning applications"', EASA. Accessed: Jan. 12, 2022. [Online]. Available: https://www.easa.europa.eu/newsroom-and-events/news/easa-releases-its-concept-paper-first-usable-guidance-level-1-machine-0

[114] H. Huang, Q. Li, and D. Zhang, 'Deep learning based image recognition for crack and leakage defects of metro shield tunnel', *Tunnelling and Underground Space Technology*, vol. 77, pp. 166–176, Jul. 2018, doi: 10.1016/j.tust.2018.04.002.

[115] D. Weimer, B. Scholz-Reiter, and M. Shpitalni, 'Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection', *CIRP Annals*, vol. 65, no. 1, pp. 417–420, Jan. 2016, doi: 10.1016/j.cirp.2016.04.072.

[116] O. Siljama, T. Koskinen, O. Jessen-Juhler, and I. Virkkunen, ‘Automated Flaw Detection in Multi-channel Phased Array Ultrasonic Data Using Machine Learning’, *J Nondestruct Eval*, vol. 40, no. 3, p. 67, Sep. 2021, doi: 10.1007/s10921-021-00796-4.

[117] I. Virkkunen, K. Miettinen, and T. Packalén, ‘Virtual flaws for NDE training and qualification’, in *e-Journal of Nondestructive Testing*, Prague, Oct. 2014, p. 8. [Online]. Available: https://www.ndt.net/?id=16779

[118] O. Janssens, R. Van de Walle, M. Loccufier, and S. Van Hoecke, ‘Deep learning for infrared thermal image based machine health monitoring’, *IEEE-ASME TRANSACTIONS ON MECHATRONICS*, vol. 23, no. 1, Art. no. 1, 2018, doi: 10.1109/TMECH.2017.2722479.

[119] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, ‘GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification’, *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018, doi: 10.1016/j.neucom.2018.09.013.

[120] S. Sundaram and N. Hulkund, ‘GAN-based Data Augmentation for Chest X-ray Classification’, Jul. 06, 2021, *arXiv*: arXiv:2107.02970. Accessed: Oct. 10, 2022. [Online]. Available: http://arxiv.org/abs/2107.02970

[121] M. M. Bejani and M. Ghatee, ‘A systematic review on overfitting control in shallow and deep neural networks’, *Artif Intell Rev*, vol. 54, no. 8, pp. 6391–6438, Dec. 2021, doi: 10.1007/s10462-021-09975-1.

[122] H. K. Dishar and L. A. Muhammed, ‘A Review of the Overfitting Problem in Convolution Neural Network and Remedy Approaches’, *J. Al-Qadisiyah Comp. Sci. Math.*, vol. 15, no. 2, Sep. 2023, doi: 10.29304/jqcm.2023.15.2.1240.

[123] ‘RollerFORM: Phased Array Wheel Probe’. Accessed: Jun. 06, 2023. [Online]. Available: https://www.olympus-ims.com/en/rollerform/

[124] ‘Ultrasonic Inspection Solutions for NDT | Peak NDT’. Accessed: Feb. 28, 2024. [Online]. Available: https://www.peakndt.com/

[125] M. Vasilev *et al.*, ‘Sensor-Enabled Multi-Robot System for Automated Welding and In-Process Ultrasonic NDE’, *Sensors*, vol. 21, no. 15, Art. no. 15, Jan. 2021, doi: 10.3390/s21155077.

[126] Y. Fu and X. Yao, ‘A review on manufacturing defects and their detection of

fiber reinforced resin matrix composites', *Composites Part C: Open Access*, vol. 8, p. 100276, Jul. 2022, doi: 10.1016/j.jcomc.2022.100276.

[127] R. Drai, F. Sellidj, M. Khelil, and A. Benchaala, 'Elaboration of some signal processing algorithms in ultrasonic techniques: application to materials NDT', *Ultrasonics*, vol. 38, no. 1, pp. 503–507, Mar. 2000, doi: 10.1016/S0041-624X(99)00082-7.

[128] A. V Oppenheim and R. W. Schafer, 'Discrete-Time Signal Processing'. Accessed: Feb. 29, 2024. [Online]. Available: https://www.pearson.com/en-us/subject-catalog/p/discrete-time-signal-processing/P200000003226/9780137549771

[129] P. Angam, K. Vijayarekha, S. Abraham, and V. Balasubramaniam, 'Fourier Analysis of Ultrasonic TOFD Signals for Defect Detection in Austenitic Stainless Steel Welds', *International Journal of Computer Applications*, vol. 71, pp. 14–17, Jun. 2013, doi: 10.5120/12385-8737.

[130] J. C. Lazaro, 'Noise reduction in ultrasonic NDT using discrete wavelet transform processing', in *2002 IEEE Ultrasonics Symposium, 2002. Proceedings.*, Oct. 2002, pp. 777–780 vol.1. doi: 10.1109/ULTSYM.2002.1193514.

[131] M. Meng, Y. J. Chua, E. Wouterson, and C. P. K. Ong, 'Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks', *Neurocomputing*, vol. 257, pp. 128–135, Sep. 2017, doi: 10.1016/j.neucom.2016.11.066.

[132] B. Wang, Y. Li, Y. Luo, X. Li, and T. Freiheit, 'Early Event Detection in a Deep-learning Driven Quality Prediction Model for Ultrasonic Welding', *Journal of Manufacturing Systems*, vol. 2021, pp. 325–336, Jun. 2021, doi: 10.1016/j.jmsy.2021.06.009.

[133] S. Lonné, L. D. Roumilly, L. L. Ber, S. Mahaut, and G. Cattiaux, 'EXPERIMENTAL VALIDATION OF CIVA ULTRASONIC SIMULATIONS', 2006, [Online]. Available: https://www.semanticscholar.org/paper/EXPERIMENTAL-VALIDATION-OF-CIVA-ULTRASONIC-Lonn%C3%A9-Roumilly/16b85af3b6a96d4657c9902ca8652fbbf93cbf2e

[134] M. Darmon *et al.*, 'VALIDATION OF AN ULTRASONIC CHARACTERIZATION TECHNIQUE FOR ANISOTROPIC MATERIALS: COMPARISON OF EXPERIMENTS WITH BEAM PROPAGATION MODELLING', presented at the 2019 INTERNATIONAL CONGRESS ON ULTRASONICS, Bruges Belgium.

[135] K. Jezzine, D. Segur, R. Ecault, and N. Dominguez, 'Simulation of ultrasonic inspections of composite structures in the CIVA software platform', *e-Journal of Nondestructive Testing*, vol. 21, no. 07, Jul. 2016, [Online]. Available: https://www.ndt.net/search/docs.php3?id=19438&msgID=0&rootID=0

[136] A. Figueira and B. Vaz, 'Survey on Synthetic Data Generation, Evaluation Methods and GANs', *Mathematics*, vol. 10, no. 15, Art. no. 15, Jan. 2022, doi: 10.3390/math10152733.

[137] I. J. Goodfellow *et al.*, 'Generative Adversarial Networks', Jun. 10, 2014, *arXiv*: arXiv:1406.2661. Accessed: Aug. 14, 2023. [Online]. Available: http://arxiv.org/abs/1406.2661

[138] A. Antoniou, A. Storkey, and H. Edwards, 'Data Augmentation Generative Adversarial Networks', *arXiv:1711.04340 [cs, stat]*, Mar. 2018, Accessed: Feb. 03, 2022. [Online]. Available: http://arxiv.org/abs/1711.04340

[139] S. Motamed, P. Rogalla, and F. Khalvati, 'Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images', *Informatics in Medicine Unlocked*, vol. 27, p. 100779, Jan. 2021, doi: 10.1016/j.imu.2021.100779.

[140] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, 'Data Synthesis based on Generative Adversarial Networks', *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018, doi: 10.14778/3231751.3231757.

[141] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, 'Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks', Aug. 24, 2020, *arXiv*: arXiv:1703.10593. Accessed: May 24, 2022. [Online]. Available: http://arxiv.org/abs/1703.10593

[142] 'EXTENDE, Experts in Non Destructive Testing Simulation with CIVA Software'. Accessed: Nov. 07, 2022. [Online]. Available: https://www.extende.com/

[143] Extende, ‘Civa 2020 User Manual’. Nov. 22, 2019.

[144] V. Nair and G. E. Hinton, ‘Rectified Linear Units Improve Restricted Boltzmann Machines’, p. 8.

[145] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, ‘Regularized Evolution for Image Classifier Architecture Search’, Feb. 16, 2019, *arXiv*: arXiv:1802.01548. doi: 10.48550/arXiv.1802.01548.

[146] ‘PyTorch’. Accessed: Jul. 26, 2023. [Online]. Available: https://www.pytorch.org

[147] C. D. Walsh, J. Edwards, and R. H. Insall, ‘Ensuring accurate stain reproduction in deep generative networks for virtual immunohistochemistry’, Apr. 14, 2022, *arXiv*: arXiv:2204.06849. Accessed: Aug. 25, 2022. [Online]. Available: http://arxiv.org/abs/2204.06849

[148] ‘scipy.stats.invgauss — SciPy v1.11.1 Manual’. Accessed: Aug. 14, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.invgauss.html

[149] ‘scipy.signal.savgol_filter — SciPy v1.11.1 Manual’. Accessed: Aug. 14, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html

[150] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[151] T. B. Hudson, P. J. Follis, J. J. Pinakidis, T. Sreekantamurthy, and F. L. Palmieri, ‘Porosity detection and localization during composite cure inside an autoclave using ultrasonic inspection’, *Composites Part A: Applied Science and Manufacturing*, vol. 147, p. 106337, Aug. 2021, doi: 10.1016/j.compositesa.2021.106337.

[152] Institute of Fundamentals of Machinery Design, Silesian University of Technology, Gliwice, 44-100, Poland, A. Wronkowicz, A. Katunin, and K. Dragan, ‘Ultrasonic C-Scan Image Processing Using Multilevel Thresholding for Damage Evaluation in Aircraft Vertical Stabilizer’, *IJIGSP*, vol. 7, no. 11, pp. 1–

8, Oct. 2015, doi: 10.5815/ijigsp.2015.11.01.

[153] Z. Zhang, M. Liu, Q. Li, and Y. Ang, 'Visualized characterization of diversified defects in thick aerospace composites using ultrasonic B-scan', *Composites Communications*, vol. 22, p. 100435, Dec. 2020, doi: 10.1016/j.coco.2020.100435.

[154] Y. Zhou *et al.*, 'Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images', *Medical Image Analysis*, vol. 70, p. 101918, May 2021, doi: 10.1016/j.media.2020.101918.

[155] Y. Liu, '3D Image Segmentation of MRI Prostate Based on a Pytorch Implementation of V-Net', *J. Phys.: Conf. Ser.*, vol. 1549, no. 4, p. 042074, Jun. 2020, doi: 10.1088/1742-6596/1549/4/042074.

[156] M. Leong, D. Prasad, Y. T. Lee, and F. Lin, 'Semi-CNN Architecture for Effective Spatio-Temporal Learning in Action Recognition', *Applied Sciences*, vol. 10, p. 557, Jan. 2020, doi: 10.3390/app10020557.

[157] A. M. Badshah *et al.*, 'Deep features-based speech emotion recognition for smart affective services', *Multimed Tools Appl*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019, doi: 10.1007/s11042-017-5292-7.

[158] H. Liu, K. Simonyan, and Y. Yang, 'DARTS: Differentiable Architecture Search', Apr. 23, 2019, *arXiv*: arXiv:1806.09055. doi: 10.48550/arXiv.1806.09055.

[159] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley, 'Neural Architecture Search without Training', Jun. 11, 2021, *arXiv*: arXiv:2006.04647. doi: 10.48550/arXiv.2006.04647.

[160] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, 'Algorithms for Hyper-Parameter Optimization', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2011. Accessed: Jun. 16, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

[161] S. Schrodi, D. Stoll, B. Ru, R. Sukthanker, T. Brox, and F. Hutter, 'Construction of Hierarchical Neural Architecture Search Spaces based on Context-free Grammars', Jun. 05, 2023, *arXiv*: arXiv:2211.01842. Accessed: Jun. 16, 2023.

[Online]. Available: http://arxiv.org/abs/2211.01842

[162] D. Maturana and S. Scherer, 'VoxNet: A 3D Convolutional Neural Network for real-time object recognition', in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 922–928. doi: 10.1109/IROS.2015.7353481.

[163] Z. Liu *et al.*, 'VB-Net: Voxel-Based Broad Learning Network for 3D Object Classification', *Applied Sciences*, vol. 10, no. 19, Art. no. 19, Jan. 2020, doi: 10.3390/app10196735.

[164] S. McKnight *et al.*, 'GANs and alternative methods of synthetic noise generation for domain adaption of defect classification of Non-destructive ultrasonic testing', Jun. 02, 2023, *arXiv*: arXiv:2306.01469. doi: 10.48550/arXiv.2306.01469.

[165] M. Gower, G. Sims, R. Lee, S. Frost, M. Stone, and M. Wall, 'Measurement Good Practice Guide', no. 78.

[166] Y. Wang, F. Tao, Y. Zuo, M. Zhang, and Q. Qi, 'Digital-Twin-Enhanced Quality Prediction for the Composite Materials', *Engineering*, vol. 22, pp. 23–33, Mar. 2023, doi: 10.1016/j.eng.2022.08.019.

[167] D. G. Puttaraju and H. G. Hanumantharaju, 'Finite element analysis and validation of tensile properties of carbon fiber reinforced polymer matrix composites', *Materials Today: Proceedings*, vol. 62, pp. 2800–2807, Jan. 2022, doi: 10.1016/j.matpr.2022.02.188.

[168] A. Hauffe, F. Hähnel, and K. Wolf, 'Comparison of algorithms to quantify the damaged area in CFRP ultrasonic scans', *Composite Structures*, vol. 235, p. 111791, Mar. 2020, doi: 10.1016/j.compstruct.2019.111791.

[169] S. Barut and N. Dominguez, 'NDT Diagnosis Automation: a Key to Efficient Production in the Aeronautic Industry', *e-Journal of Nondestructive Testing*, vol. 21, no. 07, Jul. 2016, [Online]. Available: https://www.ndt.net/search/docs.php3?id=19184&msgID=0&rootID=0

[170] S. Kumaran and S. Rani, 'Application of 6db Drop Technique to Estimate the Width of Sub Assembly Ring Top Using Pulse Echo Ultrasonic Technique', *International Journal of Engineering and Technology*, vol. 5, pp. 4771–4775, Jan. 2013.

[171] P. Ciorau, ‘Comparison Between -6 DB and -12 DB Amplitude Drop Techniques for Length Sizing’.

[172] X. Li, Y. Wang, P. Ni, H. Hu, and Y. Song, ‘Flaw sizing using ultrasonic C-scan imaging with dynamic thresholds’, *insight*, vol. 59, no. 11, pp. 603–608, Nov. 2017, doi: 10.1784/insi.2017.59.11.603.

[173] X. Cheng, G. Ma, Z. Wu, H. Zu, and X. Hu, ‘Automatic defect depth estimation for ultrasonic testing in carbon fiber reinforced composites using deep learning’, *NDT & E International*, vol. 135, p. 102804, Apr. 2023, doi: 10.1016/j.ndteint.2023.102804.

[174] P. M. Cheng and H. S. Malhi, ‘Transfer Learning with Convolutional Neural Networks for Classification of Abdominal Ultrasound Images’, *J Digit Imaging*, vol. 30, no. 2, pp. 234–243, Apr. 2017, doi: 10.1007/s10278-016-9929-2.

[175] O. Ronneberger, P. Fischer, and T. Brox, ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[176] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, ‘3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation’, Jun. 21, 2016, *arXiv*: arXiv:1606.06650. doi: 10.48550/arXiv.1606.06650.

[177] M. Sharifzadeh, H. Benali, and H. Rivaz, ‘Phase Aberration Correction: A Convolutional Neural Network Approach’, *IEEE Access*, vol. 8, pp. 162252–162260, 2020, doi: 10.1109/ACCESS.2020.3021685.

[178] K. Ono and A. Gallego, ‘Attenuation of lamb waves in CFRP plates’, *Journal of Acoustic Emission*, vol. 30, pp. 109–124, Jan. 2012.

[179] J. Yang, K. Zhou, and Z. Liu, ‘Full-Spectrum Out-of-Distribution Detection’, *Int J Comput Vis*, vol. 131, no. 10, pp. 2607–2622, Oct. 2023, doi: 10.1007/s11263-023-01811-z.

[180] S. McKnight *et al.*, ‘A comparison of methods for generating synthetic training data for domain adaption of deep learning models in ultrasonic non-destructive evaluation’, *NDT & E International*, vol. 141, p. 102978, Jan. 2024, doi:

10.1016/j.ndteint.2023.102978.

[181] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, 'Self-supervised Learning: A Succinct Review', *Arch Computat Methods Eng*, vol. 30, no. 4, pp. 2761–2775, May 2023, doi: 10.1007/s11831-023-09884-2.

[182] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, and Y. Gu, 'A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations', *Expert Systems with Applications*, vol. 242, p. 122807, May 2024, doi: 10.1016/j.eswa.2023.122807.

[183] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[184] P. P. Ray, 'ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope', *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[185] M. Oquab *et al.*, 'DINOv2: Learning Robust Visual Features without Supervision', Feb. 02, 2024, *arXiv*: arXiv:2304.07193. doi: 10.48550/arXiv.2304.07193.

[186] 'CLIP: Connecting text and images'. Accessed: Apr. 03, 2024. [Online]. Available: https://openai.com/research/clip

[187] R. He, A. Ravula, B. Kanagal, and J. Ainslie, 'RealFormer: Transformer Likes Residual Attention', Sep. 10, 2021, *arXiv*: arXiv:2012.11747. doi: 10.48550/arXiv.2012.11747.

[188] S. Gidaris, P. Singh, and N. Komodakis, 'Unsupervised Representation Learning by Predicting Image Rotations', Mar. 20, 2018, *arXiv*: arXiv:1803.07728. doi: 10.48550/arXiv.1803.07728.

[189] M. Noroozi and P. Favaro, 'Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles', Aug. 22, 2017, *arXiv*: arXiv:1603.09246. doi: 10.48550/arXiv.1603.09246.

[190] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, 'Context Encoders: Feature Learning by Inpainting', Nov. 21, 2016, *arXiv*: arXiv:1604.07379. doi: 10.48550/arXiv.1604.07379.

[191] R. Zhang, P. Isola, and A. A. Efros, ‘Colorful Image Colorization’, Oct. 05, 2016, *arXiv*: arXiv:1603.08511. doi: 10.48550/arXiv.1603.08511.

[192] G. Koch, R. Zemel, and R. Salakhutdinov, ‘Siamese Neural Networks for One-shot Image Recognition’.

[193] S. McKnight *et al.*, ‘Advancing Carbon Fiber Composite Inspection: Deep Learning-Enabled Defect Localization and Sizing via 3-Dimensional U-Net Segmentation of Ultrasonic Data’, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, pp. 1–1, 2024, doi: 10.1109/TUFFC.2024.3408314.

[194] B. Lim and S. Zohren, ‘Time-series forecasting with deep learning: a survey’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194, p. 20200209, Feb. 2021, doi: 10.1098/rsta.2020.0209.

[195] H. I. Fawaz *et al.*, ‘InceptionTime: Finding AlexNet for Time Series Classification’, *Data Min Knowl Disc*, vol. 34, no. 6, pp. 1936–1962, Nov. 2020, doi: 10.1007/s10618-020-00710-y.

[196] J. P. E. Schouten, C. Matek, L. F. P. Jacobs, M. C. Buck, D. Bošnački, and C. Marr, ‘Tens of images can suffice to train neural networks for malignant leukocyte detection’, *Sci Rep*, vol. 11, no. 1, p. 7995, Apr. 2021, doi: 10.1038/s41598-021-86995-5.

[197] ‘scikit-image: Image processing in Python — scikit-image’. Accessed: Apr. 06, 2024. [Online]. Available: https://scikit-image.org/

[198] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, ‘A Deep Siamese Convolution Neural Network for Multi-Class Classification of Alzheimer Disease’, *Brain Sci*, vol. 10, no. 2, p. 84, Feb. 2020, doi: 10.3390/brainsci10020084.

[199] ‘Aerospace Event 2024’. Accessed: Jun. 04, 2024. [Online]. Available: https://www.bindt.org/events-and-awards/PastEventsandWebinars/aerospace-event-2024/?cookie-accept=1

Appendix

Table 28: Detection accuracy across thresholds and processing steps for each sample.

Threshold	Defective Sample	Detection Accuracy % (False positives)			
		Forward Sweep	Backward Sweep	Combined Sweep	Area Threshold
0. 9999999	1	16.85 (74)	12.40 (106)	39.47 (23)	**100.00 (0)**
	2	14.79 (144)	21.55 (91)	55.56 (20)	**100.00 (0)**
	3	7.98 (173)	10.49 (128)	22.73 (51)	**100.00 (0)**
0. 999999	1	9.20 (148)	7.54 (184)	28.85 (37)	93.75 (1)
	2	8.28 (277)	13.30 (163)	42.37 (34)	100.00 (0)
	3	5.34 (266)	6.91 (202)	16.67 (75)	93.75 (1)
0. 99999	1	4.66 (307)	3.83 (377)	12.83 (102)	83.33 (3)
	2	4.28 (559)	5.94 (396)	20.49 (97)	96.15 (1)
	3	3.42 (395)	3.99 (337)	10.79 (124)	88.24 (2)
0. 9999	1	2.17 (677)	2.02 (729)	5.68 (249)	71.43 (6)
	2	2.33 (1046)	2.76 (880)	7.99 (288)	92.59 (2)
	3	2.33 (586)	2.34 (584)	5.88 (240)	78.95 (4)
0. 999	1	1.31 (1129)	1.36 (1085)	1.89 (780)	65.22 (8)
	2	1.33 (1704)	1.39 (1780)	2.32 (1053)	89.29 (3)
	3	1.68 (759)	1.71 (806)	2.85 (512)	84.62 (2)
0. 99	1	4.95 (288)	4.79 (298)	1.51 (980)	50.00 (15)
	2	5.73 (411)	3.81 (632)	1.30 (1896)	96.15 (1)
	3	4.34 (331)	3.60 (402)	1.62 (913)	60.00 (10)